MIMO Wireless Communications under Statistical Queueing Constraints

Mustafa Cenk Gursoy

Abstract

The performance of multiple-input multiple-output wireless systems is investigated in the presence of statistical queueing constraints. Queuing constraints are imposed as limitations on buffer violation probabilities. The performance under such constraints is captured through the effective capacity formulation. A detailed analysis of the effective capacity is carried out in the low-power, wideband, and high–signal-to-noise ratio (SNR) regimes. In the low-power analysis, expressions for the first and second derivatives of the effective capacity with respect to SNR at $\text{SNR} = 0$ are obtained under various assumptions on the degree of channel state information at the transmitter. Transmission strategies that are optimal in the sense of achieving the first and second derivatives are identified. It is shown that while the first derivative does not get affected by the presence of queueing constraints, the second derivative gets smaller as the constraints become more stringent. Through the energy efficiency analysis, this is shown to imply that the minimum bit energy requirements do not change with more strict limitations but the wideband slope diminishes. Similar results are obtained in the wideband regime if rich multipath fading is being experienced. On the other hand, sparse multipath fading with bounded number of degrees of freedom is shown to increase the minimum bit energy requirements in the presence of queueing constraints. Following the low-SNR study, the impact of buffer limitations on the high-SNR performance is quantified by analyzing the high-SNR slope and the power offset in Rayleigh fading channels. Finally, numerical results are provided to illustrate the theoretical findings, and to demonstrate the interactions between the queueing constraints and spatial dimensions over a wide range of SNR values.

I. Introduction

Having multiple antennas at the transmitter and receiver has been shown to improve the performance significantly in terms of both reliability and throughput when the channel fading coefficients are known at the receiver and/or transmitter. Due to these promising gains in the performance, information-theoretic analysis of multiple-input multiple-output (MIMO) channels has attracted much interest in the research

The author is with the Department of Electrical Engineering, University of Nebraska-Lincoln, Lincoln, NE, 68588 (e-mail: gursoy@engr.unl.edu).

This work was supported by the National Science Foundation under Grants CCF – 0546384 (CAREER), CNS – 0834753, and CCF-0917265. The material in this paper will be presented in part at the Forty-Seventh Annual Allerton Conference on Communication, Control, and Computing in Oct. 2009.

community. In particular, considerable effort has been expended in the study of the capacity of MIMO wireless channels (see e.g., [1] and the references therein). For instance, multiple-antenna capacity is studied in the low-power regime in [2] and [3], and in the high-SNR regime in [4]. In most studies on MIMO channel capacity, ergodic Shannon capacity formulation is employed as the main performance metric. However, this formulation does not capture the performance in the presence of quality-of-service (QoS) limitations in the form of constraints on queueing delays or queue lengths, although providing QoS assurances is of paramount importance in many delay-sensitive wireless systems, e.g., voice over IP (VoIP), and interactive and streaming video applications.

In [5], effective capacity is proposed as a metric that can be employed to measure the performance in the presence of statistical QoS limitations. Effective capacity formulation uses the large deviations theory and incorporates the statistical QoS constraints by capturing the rate of decay of the buffer occupancy probability for large queue lengths. Hence, effective capacity can be regarded as the maximum throughput of a system operating under limitations on the buffer violation probability. This formulation is tightly linked and in a sense dual to the concept of effective bandwidth [6] [7] that is employed in the analysis of how much resource in terms of service rates is needed to support a given time-varying arrival process. The analysis of the effective capacity in various wireless communication settings has been conducted in several recent studies (see e.g., [9] – [16]).

In this paper, we study the effective capacity of MIMO wireless channels. In particular, we consider the low-power, wideband, and high-SNR regimes and identify the impact of the QoS limitations[1] on the performance. We would like to note that recently references [17] and [18] have also investigated the effective capacity of multiple-antenna channels. In [17], the authors study the multiple-input single-output (MISO) channels and determine the optimal transmit strategies with covariance feedback. In [18], the concentration is on the MISO and single-input multiple-output (SIMO) channels. Analysis of MIMO channels is carried out only in the large antenna regime in which the number of receive and/or transmit antennas increase without bound. In addition, the authors in [18] consider a MIMO channel matrix with independent and identically distributed (i.i.d.) zero-mean Gaussian entries, and consider equal power allocation across the antennas. In this paper, we consider a general MIMO link model in which the fading coefficients have arbitrary distributions and are possibly correlated[2], provide a detailed study of the low-power, wideband, and high-SNR regimes, investigate the transmission strategies under various assumptions on the degree of channel knowledge at the transmitter, and identify the impact of

[1] Throughout the paper, we use the terms “QoS constraints”, “queueing constraints”, and “buffer constraints” interchangeably.

[2] Only in the high-SNR regime, we concentrate on the canonical MIMO model in which the fading coefficients are i.i.d. zero-mean, unit-variance, Gaussian random variables.

QoS constraints on the performance. The original contributions of this paper are the following:

1) We obtain expressions for the first and second derivatives of the effective capacity at $\text{SNR} = 0$ under various assumptions on the availability of channel knowledge at the transmitter, and show that while the first derivative is independent of the queueing constraints, the second derivative diminishes as the constraints become more stringent. Transmission strategies that achieve these derivatives are identified.
2) As a result of the findings on the derivatives of the effective capacity, we determine in the low-power regime that the minimum bit energy requirements in the presence of QoS limitations are the same as those attained in the absence of such constraints. On the other hand, we show that the wideband slope decreases under more strict queueing constraints, indicating that energy expenditure increases unless one is operating at the minimum bit energy level.
3) Under certain assumptions, we show that the results obtained in the low-power regime apply to the wideband regime with rich multipath fading. In contrast, we establish that sparse multipath fading has a significant impact on the performance in the wideband regime. In particular, we prove that minimum bit energies greater than that achieved in the absence of QoS constraints are required if the number of degrees of freedom in the form of noninteracting subchannels remain bounded as the bandwidth increases.
4) Considering i.i.d. Rayleigh fading channel model, we identify the effect of QoS limitations on the performance in the high-SNR regime by determining the high-SNR slope and power offset values.

The organization of the rest of the paper is as follows. We describe the MIMO channel model in Section II. In Section III, we provide a description of the effective capacity formulation, and apply it to the MIMO setting. In Section IV, we study the effective capacity in the low-power regime and determine the first and second derivatives of the effective capacity at zero SNR. Subsequently, we apply the derivative expressions to investigate the energy efficiency. In Section V, we explore the effect of QoS limitations in the wideband regime, and identify the minimum bit energy requirements. In Section VI, we concentrate on the high-SNR regime, and determine the impact of QoS constraints on the performance in the i.i.d. Rayleigh fading channel. Finally, we provide numerical results in Section VII and conclude in Section VIII.

II. Channel Model

We consider a MIMO channel model and assume that the transmitter and receiver are equipped with n_T and n_R antennas, respectively. Assuming flat-fading, we can express the channel input-output relation as

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}. \tag{1}$$

Above, $\mathbf{x}$ denotes the $n_T \times 1$–dimensional transmitted signal vector, and $\mathbf{y}$ represents the $n_R \times 1$–dimensional received signal vector. The channel input is assumed to be subject to the following average energy constraint:

$$\mathbb{E}\{\|\mathbf{x}\|^2\} \leq \frac{P}{B} \tag{2}$$

where B is the bandwidth of the system. When the bandwidth is B, we can assume that B input vectors are transmitted every second, and (2) implies that the average power of the system is limited by P. In (1), $\mathbf{n}$ with dimension $n_R \times 1$ is a zero-mean Gaussian random vector with $E\{\mathbf{n}\mathbf{n}^\dagger\} = N_0\mathbf{I}$, where $\mathbf{I}$ is the identity matrix. The signal-to-noise ratio is defined as

$$\text{SNR} = \frac{\mathbb{E}\{\|\mathbf{x}\|^2\}}{\mathbb{E}\{\|\mathbf{n}\|^2\}} = \frac{P}{n_R B N_0}. \tag{3}$$

We also define the normalized input covariance matrix as

$$\mathbf{K}_x = \frac{\mathbb{E}\{\mathbf{x}\mathbf{x}^\dagger\}}{P/B}. \tag{4}$$

Note that the average energy constraint in (2) implies that the trace of the normalized covariance matrix is upper bounded by

$$\text{tr}\,(\mathbf{K}_x) \leq 1. \tag{5}$$

Finally, in (1), $\mathbf{H}$ denotes the $n_R \times n_T$–dimensional random channel matrix whose components are the fading coefficients between the corresponding antennas at the transmitting and receiving ends. Unless specified otherwise, the components of $\mathbf{H}$ are assumed to have arbitrary distributions with finite variances. Additionally, we consider the block-fading scenario and assume that the realization of the matrix $\mathbf{H}$ remains fixed over a block of duration T seconds and changes independently from one block to another.

III. Effective Capacity of a MIMO Link

In [5], Wu and Negi defined the effective capacity as the maximum constant arrival rate that a given service process can support in order to guarantee a statistical QoS requirement specified by the QoS exponent θ [3]. If we define Q as the stationary queue length, then θ is the decay rate of the tail of the distribution of the queue length Q:

$$\lim_{q \to \infty} \frac{\log P(Q \geq q)}{q} = -\theta. \tag{6}$$

Therefore, for large $q_{\max}$, we have the following approximation for the buffer violation probability: $P(Q \geq q_{\max}) \approx e^{-\theta q_{\max}}$. Hence, while larger θ corresponds to more strict QoS constraints, smaller θ implies looser QoS guarantees. Similarly, if D denotes the steady-state delay experienced in the buffer, then $P(D \geq d_{\max}) \approx e^{-\theta \delta d_{\max}}$ for large $d_{\max}$, where δ is determined by the arrival and service processes [11]. Therefore, effective capacity formulation provides the maximum constant arrival rates that can be supported by the time-varying wireless channel under the queue length constraint $P(Q \geq q_{\max}) \leq e^{-\theta q_{max}}$ for large q_{max} or the delay constraint $P(D \geq d_{\max}) \leq e^{-\theta \delta d_{\max}}$ for large $d_{\max}$. Since the average arrival rate is equal to the average departure rate when the queue is in steady-state [8], effective capacity can also be seen as the maximum throughput in the presence of such constraints.

The effective capacity is given by ([5], [6], [7])

$$-\frac{\Lambda(-\theta)}{\theta} = -\lim_{t \to \infty} \frac{1}{\theta t} \log_e \mathbb{E}\{e^{-\theta S[t]}\} \tag{7}$$

where $S[t] = \sum_{i=1}^{t} R[i]$ is the time-accumulated service process and $\{R[i], i = 1, 2, \ldots\}$ denotes the discrete-time stationary and ergodic stochastic service process. Under the block-fading assumption, the effective capacity formulation simplifies to

$$-\frac{\Lambda(-\theta)}{\theta} = -\frac{1}{\theta T} \log_e \mathbb{E}\{e^{-\theta T R[i]}\}. \tag{8}$$

Under a short-term power constraint, the stochastic service process in a MIMO channel with a given normalized input covariance matrix $\mathbf{K}_x$ is

$$B \log_2 \det \left(\mathbf{I} + \frac{P}{B N_0} \mathbf{H} \mathbf{K}_x \mathbf{H}^\dagger \right) = B \log_2 \det \left(\mathbf{I} + n_R \text{SNR} \mathbf{H} \mathbf{K}_x \mathbf{H}^\dagger \right) \text{ bits/s} \tag{9}$$

where B denotes the bandwidth of the system and SNR is as defined in (3). We first consider the case in which $\mathbf{H}$ is perfectly-known at the receiver and transmitter. In this scenario, the transmitter can adapt the input covariance matrix with respect to each realization of $\mathbf{H}$ in order to maximize the service rate. Therefore, using the formulation in (8), we can express the effective capacity normalized by the

[3] For time-varying arrival rates, effective capacity specifies the effective bandwidth of the arrival process that can be supported by the channel.

bandwidth and the receive dimensions as

$$\mathsf{C}_E(\text{SNR},\theta) = -\frac{1}{\theta T B n_R}\log_e \mathbb{E}\left\{\exp\left(-\theta T B \max_{\substack{\mathbf{K}_x \succeq \mathbf{0}\\ \text{tr}\,(\mathbf{K}_x)\leq 1}} \log_2 \det\left(\mathbf{I} + n_R \text{SNR}\mathbf{H}\mathbf{K}_x\mathbf{H}^\dagger\right)\right)\right\} \text{ bits/s/Hz/dimension} \tag{10}$$

As θ vanishes, the QoS constraints become loose and it can be easily verified that the effective capacity approaches the ergodic channel capacity, i.e.,

$$\lim_{\theta\to 0}\mathsf{C}_E(\text{SNR},\theta) = \frac{1}{n_R}\mathbb{E}\left\{\max_{\substack{\mathbf{K}_x \succeq \mathbf{0}\\ \text{tr}\,(\mathbf{K}_x)\leq 1}} \log_2 \det\left(\mathbf{I} + n_R \text{SNR}\mathbf{H}\mathbf{K}_x\mathbf{H}^\dagger\right)\right\}. \tag{11}$$

For $\theta > 0$, the effective capacity is in general smaller than the ergodic capacity. We can easily see this by interchanging the logarithm and the expectation in (10) and applying the Jensen's inequality:

$$\mathsf{C}_E(\text{SNR},\theta) = -\frac{1}{\theta T B n_R}\log_e \mathbb{E}\left\{\exp\left(-\theta T B \max_{\substack{\mathbf{K}_x \succeq \mathbf{0}\\ \text{tr}\,(\mathbf{K}_x)\leq 1}} \log_2 \det\left(\mathbf{I} + n_R \text{SNR}\mathbf{H}\mathbf{K}_x\mathbf{H}^\dagger\right)\right)\right\} \tag{12}$$

$$\leq -\frac{1}{\theta T B n_R}\mathbb{E}\left\{\log_e \exp\left(-\theta T B \max_{\substack{\mathbf{K}_x \succeq \mathbf{0}\\ \text{tr}\,(\mathbf{K}_x)\leq 1}} \log_2 \det\left(\mathbf{I} + n_R \text{SNR}\mathbf{H}\mathbf{K}_x\mathbf{H}^\dagger\right)\right)\right\} \tag{13}$$

$$= \frac{1}{n_R}\mathbb{E}\left\{\max_{\substack{\mathbf{K}_x \succeq \mathbf{0}\\ \text{tr}\,(\mathbf{K}_x)\leq 1}} \log_2 \det\left(\mathbf{I} + n_R \text{SNR}\mathbf{H}\mathbf{K}_x\mathbf{H}^\dagger\right)\right\}. \tag{14}$$

Above, we have assumed that $\mathbf{H}$ is perfectly known at the transmitter. If, on the other hand, only statistical information regarding $\mathbf{H}$ is available at the transmitter, then the input covariance matrix can be chosen to maximize the effective capacity. In such a case, the normalized effective capacity can be expressed as

$$\mathsf{C}_E(\text{SNR},\theta) = \max_{\substack{\mathbf{K}_x \succeq \mathbf{0}\\ \text{tr}\,(\mathbf{K}_x)\leq 1}} -\frac{1}{\theta T B n_R}\log_e \mathbb{E}\left\{\exp\left(-\theta T B \log_2 \det\left(\mathbf{I} + n_R \text{SNR}\mathbf{H}\mathbf{K}_x\mathbf{H}^\dagger\right)\right)\right\} \text{ bits/s/Hz/dimension.} \tag{15}$$

For a given (and not necessarily optimal) input covariance matrix $\mathbf{K}_x$, we call the throughput as effective rate and express it as

$$\mathsf{R}_E(\text{SNR},\theta) = -\frac{1}{\theta T B n_R}\log_e \mathbb{E}\left\{\exp\left(-\theta T B \log_2 \det\left(\mathbf{I} + n_R \text{SNR}\mathbf{H}\mathbf{K}_x\mathbf{H}^\dagger\right)\right)\right\} \text{ bits/s/Hz/dimension.} \tag{16}$$

In practice, uniform power allocation across the antennas might be preferred. In this case, $\mathbf{K}_x = \frac{1}{n_T}\mathbf{I}$, and the effective rate can be written as

$$\mathsf{R}_{E,\text{id}}(\text{SNR},\theta) = -\frac{1}{\theta T B n_R}\log_e \mathbb{E}\left\{\exp\left(-\theta T B \log_2 \det\left(\mathbf{I} + \frac{n_R}{n_T}\text{SNR}\mathbf{H}\mathbf{H}^\dagger\right)\right)\right\} \text{ bits/s/Hz/dimension} \tag{17}$$

where the subscript “id” is introduced to denote that this expression is the throughput when the covariance matrix is proportional to an identity matrix.

Note that the effective capacity and effective rate expressions in (10), (15), (16), and (17) are proportional to the logarithm of the moment generating function of the instantaneous transmission rates.

Since the subsequent analysis assumes that the QoS exponent is fixed as power diminishes or increases or bandwidth increases, we generally suppress the argument θ and write the effective capacity and rate as $\mathsf{C}_E(\text{SNR})$ and $\mathsf{R}_E(\text{SNR})$, respectively.

Finally, before we go through a more detailed analysis of the effective capacity in the following sections, we would like to discuss several implicit assumptions made in the formulations provided in this section. The service rate expression in (9) implies that the maximum transmission rates are equal to the instantaneous channel capacity in each block of duration T. Hence, we implicitly assume that the number of symbols in each block, TB, is large enough for this assumption to have operational meaning in practice. In (15), it is assumed that the service rate is still given by (9) and hence the transmitter employs variable-rate transmission scheme, even though the transmitter does not know the instantaneous realizations of $\mathbf{H}$. Note this can be accomplished by using recently developed rateless codes such as LT [19] or Raptor [20] codes, which enable the transmitter to adapt its rate to the channel realization without requiring CSI at the transmitter side [21], [22]. It is also important to note that the analysis conducted in this paper apply in the large-queue-length regime. If the buffer size is finite and small, then the arrival rates that can be supported by the system will be smaller than those considered in the paper, and in this case, one has to consider packet loss probabilities as well. Therefore, if the above-mentioned conditions and assumptions are not satisfied in the system, then the performance degradation will be more severe. For such cases, the results of this paper can be seen as fundamental limits (or upper bounds) which can serve as benchmarks for system performance.

IV. Effective Capacity in the Low-Power Regime

A. First and Second Derivatives of the Effective Capacity

In this section, we study the effective capacity in the low-SNR regime and investigate the impact of the QoS exponent θ. In particular, we consider the following second-order expansion of the effective capacity under different assumptions on the degree of channel state information:

$$\mathsf{C}_E(\text{SNR}) = \dot{\mathsf{C}}_E(0)\text{SNR} + \ddot{\mathsf{C}}_E(0)\frac{\text{SNR}^2}{2} + o(\text{SNR}^2) \tag{18}$$

where $\dot{\mathsf{C}}_E(0)$ and $\ddot{\mathsf{C}}_E(0)$ denote the first and second derivatives of the effective capacity with respect to SNR at $\text{SNR} = 0$. We first have the following result when the channel is perfectly known at the transmitter and receiver.

Theorem 1: Assume that the realizations of the channel matrix $\mathbf{H}$ are perfectly known at the receiver and transmitter. Assume further that the transmitter is subject to a short-term power constraint and hence is not allowed to perform power adaptation over time. Then, the first and second derivatives of the effective capacity in (10) with respect to SNR at $\text{SNR} = 0$ are

$$\dot{\mathsf{C}}_E(0) = \frac{1}{\log_e 2}\mathbb{E}\{\lambda_{\max}(\mathbf{H}^\dagger\mathbf{H})\} \tag{19}$$

and

$$\ddot{\mathsf{C}}_E(0) = \frac{\theta T B n_R}{\log_e^2 2}\left[\mathbb{E}^2\{\lambda_{\max}(\mathbf{H}^\dagger\mathbf{H})\} - \mathbb{E}\{\lambda_{\max}^2(\mathbf{H}^\dagger\mathbf{H})\}\right] - \frac{n_R}{l\log_e 2}\mathbb{E}\{\lambda_{\max}^2(\mathbf{H}^\dagger\mathbf{H})\} \tag{20}$$

where $\lambda_{\max}(\mathbf{H}^\dagger\mathbf{H})$ denotes the maximum eigenvalue of $\mathbf{H}^\dagger\mathbf{H}$, and l is the multiplicity of $\lambda_{\max}(\mathbf{H}^\dagger\mathbf{H})$.

Proof: For a given input covariance matrix $\mathbf{K}_x$, the effective rate is expressed as

$$\mathsf{R}_E(\text{SNR}) = -\frac{1}{\theta T B n_R}\log_e \mathbb{E}\left\{\exp\left(-\theta T B \log_2 \det\left(\mathbf{I} + n_R \text{SNR}\mathbf{H}\mathbf{K}_x\mathbf{H}^\dagger\right)\right)\right\} \tag{21}$$

$$= -\frac{1}{\theta T B n_R}\log_e \mathbb{E}\left\{\exp\left(-\theta T B \log_2 \det\left(\mathbf{I} + n_R \text{SNR}\mathbf{\Phi}\right)\right)\right\} \tag{22}$$

$$= -\frac{1}{\theta T B n_R}\log_e \mathbb{E}\left\{\exp\left(-\theta T B \sum_i \log_2\left(1 + n_R \text{SNR}\lambda_i(\mathbf{\Phi})\right)\right)\right\} \tag{23}$$

$$= -\frac{1}{\theta T B n_R}\log_e \mathbb{E}\left\{\exp\left(-\frac{\theta T B}{\log_e 2} \sum_i \log_e\left(1 + n_R \text{SNR}\lambda_i(\mathbf{\Phi})\right)\right)\right\} \tag{24}$$

$$= -\frac{1}{\theta T B n_R}\log_e \mathbb{E}\left\{f(\text{SNR}, \theta)\right\}. \tag{25}$$

In (22) above, we have defined $\mathbf{\Phi} = \mathbf{H}\mathbf{K}_x\mathbf{H}^\dagger$. (23) is obtained by noting that the determinant of a matrix is equal to the product of its eigenvalues, i.e., $\det(\mathbf{I} + n_R\text{SNR}\mathbf{\Phi}) = \prod_i(1 + n_R\text{SNR}\lambda_i(\mathbf{\Phi}))$, and also using the fact that the logarithm of a product is equal to the sum of the logarithms of the terms in the product. In (24), the base of the logarithm is changed from 2 to e. In (25), we have defined the function $f(\text{SNR}, \theta) = \exp\left(-\frac{\theta TB}{\log_e 2}\sum_i \log_e(1 + n_R\text{SNR}\lambda_i(\mathbf{\Phi}))\right)$.

Now, taking the derivative of R_E with respect to SNR yields

$$\dot{\mathsf{R}}_E(\text{SNR}) = -\frac{1}{\theta TBn_R}\frac{1}{\mathbb{E}\{f(\text{SNR},\theta)\}}\mathbb{E}\left\{-\frac{\theta TB}{\log_e 2}\sum_i \frac{n_r\lambda_i(\mathbf{\Phi})}{1 + n_r\text{SNR}\lambda_i(\mathbf{\Phi})}f(\text{SNR},\theta)\right\}. \tag{26}$$

Noting that the function f evaluated at $\text{SNR} = 0$ is one, i.e., $f(0, \theta) = 1$, we can easily see from (26) that the value of the first derivative of the effective rate at $\text{SNR} = 0$ is

$$\dot{\mathsf{R}}_E(0) = \frac{1}{\log_e 2}\mathbb{E}\left\{\sum_i \lambda_i(\mathbf{\Phi})\right\} = \frac{1}{\log_e 2}\mathbb{E}\left\{\text{tr}\left(\mathbf{\Phi}\right)\right\} = \frac{1}{\log_e 2}\mathbb{E}\left\{\text{tr}\left(\mathbf{H}\mathbf{K}_x\mathbf{H}^\dagger\right)\right\} \tag{27}$$

where we have used the fact that the sum of the eigenvalues of a matrix is equal to its trace. Note that the normalized input covariance matrix $\mathbf{K}_x$ is by definition a positive semidefinite Hermitian matrix. As a Hermitian matrix, $\mathbf{K}_x$ can be written as [31, Theorem 4.1.5]

$$\mathbf{K}_x = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\dagger = \sum_{i=1}^{n_T} d_i\mathbf{u}_i\mathbf{u}_i^\dagger \tag{28}$$

where $\mathbf{U}$ is a unitary matrix, $\{\mathbf{u}_i\}$ are the column vectors of $\mathbf{U}$ and form an orthonormal set, $\mathbf{\Lambda}$ is a real diagonal matrix, $\{d_i\}$ are the diagonal components of $\mathbf{\Lambda}$. Since $\mathbf{K}_x$ is positive semidefinite, we have $d_i \geq 0$. Moreover, since all available energy should be used for transmission (i.e., the average energy and hence trace constraints should be satisfied with equality), we have $\text{tr}\left(\mathbf{K}_x\right) = \sum_{i=1}^{n_T} d_i = 1$. Combining (27) and (28), we can now write

$$\dot{\mathsf{R}}_E(0) = \frac{1}{\log_e 2}\mathbb{E}\left\{\text{tr}\left(\mathbf{H}\mathbf{K}_x\mathbf{H}^\dagger\right)\right\} = \frac{1}{\log_e 2}\sum_{i=1}^{n_T} d_i\mathbb{E}\left\{\text{tr}\left(\mathbf{H}\mathbf{u}_i\mathbf{u}_i^\dagger\mathbf{H}^\dagger\right)\right\} \tag{29}$$

$$= \frac{1}{\log_e 2}\sum_{i=1}^{n_T} d_i\mathbb{E}\left\{\mathbf{u}_i^\dagger\mathbf{H}^\dagger\mathbf{H}\mathbf{u}_i\right\} \tag{30}$$

$$\leq \frac{1}{\log_e 2}\mathbb{E}\{\lambda_{\max}(\mathbf{H}^\dagger\mathbf{H})\}. \tag{31}$$

where $\lambda_{\max}(\mathbf{H}^\dagger\mathbf{H})$ denotes the maximum eigenvalue of the matrix $\mathbf{H}^\dagger\mathbf{H}$. The upper bound in (31) follows from the facts that $d_i \in [0, 1]$ and $\sum_i d_i = 1$, and from [31, Theorem 4.2.2] which states that since $\mathbf{H}^\dagger\mathbf{H}$ is a Hermitian matrix and $\{\mathbf{u}_i\}$ are unit vectors, we have

$$\mathbf{u}_i^\dagger\mathbf{H}^\dagger\mathbf{H}\mathbf{u}_i \leq \lambda_{\max}(\mathbf{H}^\dagger\mathbf{H}) \quad \forall i. \tag{32}$$

The upper bound in (31) can be achieved by beamforming in the direction in which $\lambda_{\max}(\mathbf{H}^\dagger\mathbf{H})$ is achieved, i.e., by choosing the normalized input covariance matrix as

$$\mathbf{K}_x = \mathbf{u}\mathbf{u}^\dagger \tag{33}$$

where $\mathbf{u}$ is the unit-norm eigenvector that corresponds to the maximum eigenvalue $\lambda_{\max}(\mathbf{H}^\dagger\mathbf{H})$. This lets us conclude that

$$\dot{\mathsf{C}}_E(0) = \frac{1}{\log_e 2}\mathbb{E}\{\lambda_{\max}(\mathbf{H}^\dagger\mathbf{H})\} \tag{34}$$

proving (19).

Before proceeding to the proof of the second derivative result, we would like to note that transmission in the maximal-eigenvalue eigenspace of $\mathbf{H}^\dagger\mathbf{H}$ is indeed necessary to achieve the first derivative. Therefore, it is also necessary to attain the second derivative of the effective capacity at zero SNR. In a general scenario in which $\lambda_{\max}(\mathbf{H}^\dagger\mathbf{H})$ has a multiplicity of $l \geq 1$, an input covariance matrix in the following form is required:

$$\mathbf{K}_x = \sum_{i=1}^{l} \alpha_i \mathbf{u}_i \mathbf{u}_i^\dagger \tag{35}$$

where $\alpha_i \in [0,1]$ and $\sum_{i=1}^{l} \alpha_i = 1$, and $\{\mathbf{u}_i\}_{i=1}^{l}$ are the orthonormal eigenvectors that span the maximal-eigenvalue eigenspace of $\mathbf{H}^\dagger\mathbf{H}$.

Now, we turn to the analysis of the second derivative. Differentiating $\dot{\mathsf{R}}_E$ in (26) once more with respect to SNR, we obtain

$$\begin{aligned}\ddot{\mathsf{R}}_E(\text{SNR}) = &-\frac{1}{\log_e 2}\frac{\mathbb{E}\left\{-\frac{\theta TB}{\log_e 2}\sum_i \frac{n_r\lambda_i(\mathbf{\Phi})}{1+n_r\text{SNR}\lambda_i(\mathbf{\Phi})} f(\text{SNR},\theta)\right\}}{\mathbb{E}^2\left\{f(\text{SNR},\theta)\right\}}\mathbb{E}\left\{\sum_i \frac{\lambda_i(\mathbf{\Phi})}{1+n_r\text{SNR}\lambda_i(\mathbf{\Phi})} f(\text{SNR},\theta)\right\} \\ &+\frac{1}{\log_e 2}\frac{1}{\mathbb{E}\left\{f(\text{SNR},\theta)\right\}}\mathbb{E}\left\{\sum_i \frac{-n_r\lambda_i^2(\mathbf{\Phi})}{(1+n_r\text{SNR}\lambda_i(\mathbf{\Phi}))^2} f(\text{SNR},\theta)\right\} \\ &-\frac{\theta TBn_R}{\log_e^2 2}\frac{1}{\mathbb{E}\left\{f(\text{SNR},\theta)\right\}}\mathbb{E}\left\{\left(\sum_i \frac{\lambda_i(\mathbf{\Phi})}{1+n_r\text{SNR}\lambda_i(\mathbf{\Phi})}\right)^2 f(\text{SNR},\theta)\right\}.\end{aligned} \tag{36}$$

Again noting that $f(0,\theta)=1$, we have

$$\ddot{\mathsf{R}}_E(0)=\frac{\theta TBn_R}{\log_e^2 2}\left(\mathbb{E}^2\left\{\sum_i \lambda_i(\mathbf{\Phi})\right\}-\mathbb{E}\left\{\left(\sum_i \lambda_i(\mathbf{\Phi})\right)^2\right\}\right)-\frac{n_R}{\log_e 2}\mathbb{E}\left\{\sum_i \lambda_i^2(\mathbf{\Phi})\right\} \tag{37}$$

$$=\frac{\theta TBn_R}{\log_e^2 2}\left(\mathbb{E}^2\left\{\text{tr}\,(\mathbf{\Phi})\right\}-\mathbb{E}\left\{\text{tr}^{\,2}(\mathbf{\Phi})\right\}\right)-\frac{n_R}{\log_e 2}\mathbb{E}\left\{\text{tr}\,(\mathbf{\Phi}^\dagger\mathbf{\Phi})\right\}. \tag{38}$$

In obtaining (38), we have used the facts that $\sum_i \lambda_i(\mathbf{\Phi})=\text{tr}\,(\mathbf{\Phi})$ and $\sum_i \lambda_i^2(\mathbf{\Phi})=\text{tr}\,(\mathbf{\Phi}^\dagger\mathbf{\Phi})$.

As described above, an input covariance matrix that is in the form given in (35) is required to achieve the second derivative of the effective capacity at $\text{SNR}=0$. For such a covariance matrix, it can be easily verified that

$$\mathbb{E}\left\{\text{tr}\,(\mathbf{\Phi})\right\}=\mathbb{E}\left\{\text{tr}\,(\mathbf{H}\mathbf{K}_x\mathbf{H}^\dagger)\right\}=\mathbb{E}\left\{\lambda_{\max}(\mathbf{H}^\dagger\mathbf{H})\right\} \tag{39}$$

and

$$\mathbb{E}\left\{\text{tr}\,(\mathbf{\Phi}^\dagger\mathbf{\Phi})\right\}=\mathbb{E}\left\{\text{tr}\,(\mathbf{H}\mathbf{K}_x\mathbf{H}^\dagger\mathbf{H}\mathbf{K}_x\mathbf{H}^\dagger)\right\}=\mathbb{E}\left\{\sum_{i,j}^{l}\alpha_i\alpha_j|\mathbf{u}_j^\dagger\mathbf{H}^\dagger\mathbf{H}\mathbf{u}_i|^2\right\} \tag{40}$$

$$=\mathbb{E}\left\{\lambda_{\max}^2(\mathbf{H}^\dagger\mathbf{H})\sum_{i,j}^{l}\alpha_i\alpha_j|\mathbf{u}_j^\dagger\mathbf{u}_i|^2\right\} \tag{41}$$

$$=\mathbb{E}\left\{\lambda_{\max}^2(\mathbf{H}^\dagger\mathbf{H})\sum_{i=1}^{l}\alpha_i^2\right\} \tag{42}$$

$$\geq\frac{1}{l}\mathbb{E}\left\{\lambda_{\max}^2(\mathbf{H}^\dagger\mathbf{H})\right\} \tag{43}$$

where (41) follows from the fact that $\{\mathbf{u}_i\}$ are the eigenvectors that correspond to $\lambda_{\max}(\mathbf{H}^\dagger\mathbf{H})$ and hence $\mathbf{H}^\dagger\mathbf{H}\mathbf{u}_i=\lambda_{\max}(\mathbf{H}^\dagger\mathbf{H})\mathbf{u}_i$, (42) follows from the orthonormality of $\{\mathbf{u}_i\}$ which implies that

$$\mathbf{u}_j^\dagger\mathbf{u}_i=\begin{cases}1 & \text{if } i=j\\ 0 & \text{if } i\neq j\end{cases}. \tag{44}$$

Finally, (43) follows from the properties that $\alpha_i\in[0,1]$ and $\sum_{i=1}^l\alpha_i=1$, and the fact that $\sum_{i=1}^l\alpha_i^2$ under these properties is minimized by choosing $\alpha_i=\frac{1}{l}$, which leads to the lower bound $\sum_{i=1}^l\alpha_i^2\geq\frac{1}{l}$.

We note from (39) that given the required the covariance structure in (35), the first term in the expression of $\ddot{\mathsf{R}}_E(0)$ in (38) is $\frac{\theta TBn_R}{\log_e^2 2}\left(\mathbb{E}^2\left\{\text{tr}\,(\mathbf{\Phi})\right\}-\mathbb{E}\left\{\text{tr}^{\,2}(\mathbf{\Phi})\right\}\right)=\frac{\theta TBn_R}{\log_e^2 2}\left(\mathbb{E}^2\left\{\lambda_{\max}(\mathbf{H}^\dagger\mathbf{H})\right\}-\mathbb{E}\left\{\lambda_{\max}^2(\mathbf{H}^\dagger\mathbf{H})\right\}\right)$ for all possible $\{\alpha_i\}$. On the other hand, the second term in (38) is minimized by having $\alpha_i=\frac{1}{l}$ for all i, i.e., by equally allocating the power in the orthogonal directions in the maximal-eigenvalue eigenspace.

Therefore, the input covariance matrix $\mathbf{K}_x = \frac{1}{l}\sum_{i=1}^{l}\mathbf{u}_i\mathbf{u}_i^\dagger$ maximizes $\ddot{\mathsf{R}}_E(0)$, and we have

$$\ddot{\mathsf{C}}_E(0) = \frac{\theta T B n_R}{\log_e^2 2}\left(\mathbb{E}^2\left\{\lambda_{\max}(\mathbf{H}^\dagger\mathbf{H})\right\} - \mathbb{E}\left\{\lambda_{\max}^2(\mathbf{H}^\dagger\mathbf{H})\right\}\right) - \frac{n_R}{l\log_e 2}\mathbb{E}\left\{\lambda_{\max}^2(\mathbf{H}^\dagger\mathbf{H})\right\} \tag{45}$$

proving (20). ■

Next, we consider the case in which the transmitter has only statistical knowledge of the channel.

Theorem 2: Assume that while the receiver perfectly knows the channel matrix $\mathbf{H}$, the transmitter only has the knowledge of $\mathbb{E}\{\mathbf{H}^\dagger\mathbf{H}\}$. Then, the first and second derivatives of the effective capacity in (15) are

$$\dot{\mathsf{C}}_E(0) = \frac{1}{\log_e 2}\lambda_{\max}(\mathbb{E}\{\mathbf{H}^\dagger\mathbf{H}\}) \tag{46}$$

and

$$\ddot{\mathsf{C}}_E(0) = \frac{\theta T B n_R}{\log_e^2 2}\lambda_{\max}^2(\mathbb{E}\{\mathbf{H}^\dagger\mathbf{H}\}) - \min_{\substack{\{\alpha_i\} \\ \alpha_i\in[0,1]\,\forall i \\ \sum_{i=1}^{l}\alpha_i=1}} \sum_{i,j}^{l}\alpha_i\alpha_j\left(\frac{\theta T B n_R}{\log_e^2 2}\mathbb{E}\{(\mathbf{u}_i^\dagger\mathbf{H}^\dagger\mathbf{H}\mathbf{u}_i)(\mathbf{u}_j^\dagger\mathbf{H}^\dagger\mathbf{H}\mathbf{u}_j)\} + \frac{n_R}{\log_e 2}\mathbb{E}\{|\mathbf{u}_j^\dagger\mathbf{H}^\dagger\mathbf{H}\mathbf{u}_i|^2\}\right) \tag{47}$$

where $\lambda_{\max}(\mathbb{E}\{\mathbf{H}^\dagger\mathbf{H}\})$ denotes the maximum eigenvalue of $\mathbb{E}\{\mathbf{H}^\dagger\mathbf{H}\}$, and l is the multiplicity of $\lambda_{\max}(E\{\mathbf{H}^\dagger\mathbf{H}\})$.

Proof: Note from (30) that for a given covariance matrix $\mathbf{K}_x = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\dagger = \sum_{i=1}^{n_T} d_i\mathbf{u}_i\mathbf{u}_i^\dagger$, the first derivative of the effective rate is

$$\begin{aligned}
\dot{\mathsf{R}}_E(0) &= \frac{1}{\log_e 2}\sum_{i=1}^{n_T} d_i\mathbb{E}\left\{\mathbf{u}_i^\dagger\mathbf{H}^\dagger\mathbf{H}\mathbf{u}_i\right\} && (48)\\
&= \frac{1}{\log_e 2}\sum_{i=1}^{n_T} d_i\mathbf{u}_i^\dagger\mathbb{E}\left\{\mathbf{H}^\dagger\mathbf{H}\right\}\mathbf{u}_i && (49)\\
&\leq \frac{1}{\log_e 2}\lambda_{\max}(\mathbb{E}\{\mathbf{H}^\dagger\mathbf{H}\}) && (50)
\end{aligned}$$

where (49) follows by noting that the transmitter has only statistical knowledge of $\mathbf{H}$, and the input covariance matrix and hence $\{\mathbf{u}_i\}$ cannot depend on the realizations of $\mathbf{H}$. Therefore, $\{\mathbf{u}_i\}$ are deterministic and can be taken out of the expectation. Now, the upper bound in (50), similarly as discussed in the proof of Theorem 1, is achieved by transmitting in the maximal-eigenvalue eigenspace of $\mathbb{E}\{\mathbf{H}^\dagger\mathbf{H}\}$. Therefore, a covariance matrix in the following form is required to achieve the first derivative of the effective capacity:

$$\mathbf{K}_x = \sum_{i=1}^{l} \alpha_i \mathbf{u}_i \mathbf{u}_i^\dagger \tag{51}$$

where $\{\mathbf{u}_i\}$ are the orthonormal eigenvectors spanning the maximal-eigenvalue eigenspace of $\mathbb{E}\{\mathbf{H}^\dagger\mathbf{H}\}$, l is the multiplicity of $\lambda_{\max}(\mathbb{E}\{\mathbf{H}^\dagger\mathbf{H}\})$, and $\{\alpha_i\}$ are constants taking values in $[0,1]$ and has unit sum, i.e., $\sum_{i=1}^{l}\alpha_i = 1$. Consequently, this covariance structure is also necessary to attain the second derivative of the effective capacity. Employing the second derivative expression in (38) with the covariance matrix in (51), and maximizing $\dot{\mathsf{R}}_E(0)$ with respect to all possible choices of $\{\alpha_i\}$, we easily obtain (47). ■

Using the results seen in the proofs of Theorems 1 and 2, we can also immediately obtain the following result when the power is uniformly distributed across the transmit antennas and hence we have $\mathbf{K}_x = \frac{1}{n_T}\mathbf{I}$.

Corollary 1: Assume that the input covariance matrix is $\mathbf{K}_x = \frac{1}{n_T}\mathbf{I}$. Then, the first and second derivatives of the effective rate $\mathsf{R}_{E,\text{id}}$ given in (17) are

$$\dot{\mathsf{R}}_{E,\text{id}}(0) = \frac{1}{n_T \log_e 2}\mathbb{E}\{\text{tr}\,(\mathbf{H}^\dagger\mathbf{H})\} \tag{52}$$

and

$$\ddot{\mathsf{R}}_{E,\text{id}}(0) = \frac{\theta T B n_R}{n_T^2 \log_e^2 2}\left[\mathbb{E}^2\{\text{tr}\,(\mathbf{H}^\dagger\mathbf{H})\} - \mathbb{E}\{\text{tr}^2(\mathbf{H}^\dagger\mathbf{H})\}\right] - \frac{n_R}{n_T^2 \log_e 2}\mathbb{E}\{\text{tr}\,((\mathbf{H}^\dagger\mathbf{H})^2)\}. \tag{53}$$

Remark 1: Note that the common theme in the results of Theorems 1 and 2, and Corollary 1 is that the first derivative does not depend on θ and hence does not get affected by the presence of QoS constraints. Indeed, the first derivative expressions are equal to the ones obtained when Shannon capacity, rather than effective capacity, is considered [2]. On the other hand, the second derivative is a function of θ and in general decreases as θ increases or equivalently as the queueing constraints become more stringent[4].

B. Energy Efficiency in the Low-Power Regime

The expressions of the first and second derivatives enable us to analyze the energy efficiency in the low-power regime. The minimum bit energy under QoS constraints is given by [2]

$$\frac{E_b}{N_0}_{\min} = \lim_{\text{SNR}\to 0}\frac{\text{SNR}}{\mathsf{C}_E(\text{SNR})} = \frac{1}{\dot{\mathsf{C}}_E(0)}. \tag{54}$$

[4] Note that $\mathbb{E}^2\{\lambda_{\max}(\mathbf{H}^\dagger\mathbf{H})\} \le \mathbb{E}\{\lambda_{\max}^2(\mathbf{H}^\dagger\mathbf{H})\}$ and $\mathbb{E}^2\{\text{tr}\,(\mathbf{H}^\dagger\mathbf{H})\} \le \mathbb{E}\{\text{tr}^2(\mathbf{H}^\dagger\mathbf{H})\}$.

At $\frac{E_b}{N_0}_{\min}$, the slope $\mathcal{S}_0$ of the spectral efficiency versus E_b/N_0 (in dB) curve is defined as [2]

$$\mathcal{S}_0 = \lim_{\frac{E_b}{N_0} \downarrow \frac{E_b}{N_0}_{\min}} \frac{\mathsf{C}_E(\frac{E_b}{N_0})}{10\log_{10}\frac{E_b}{N_0} - 10\log_{10}\frac{E_b}{N_0}_{\min}} 10\log_{10} 2. \tag{55}$$

Considering the expression for normalized effective capacity, the wideband slope can be found from [2][5]

$$\mathcal{S}_0 = \frac{2(\dot{\mathsf{C}}_E(0))^2}{-\ddot{\mathsf{C}}_E(0)} \log_e 2 \quad \text{bits/s/Hz/(3 dB)/receive antenna.} \tag{56}$$

Corollary 2: Applying the results of Theorem 1 to the above formulation, we obtain

$$\frac{E_b}{N_0}_{\min} = \frac{\log_e 2}{\mathbb{E}\{\lambda_{\max}(\mathbf{H}^\dagger\mathbf{H})\}} \tag{57}$$

$$\begin{aligned} \mathcal{S}_0 &= \frac{2\mathbb{E}^2\{\lambda_{\max}(\mathbf{H}^\dagger\mathbf{H})\}}{\frac{n_R}{l}\mathbb{E}\{\lambda^2_{\max}(\mathbf{H}^\dagger\mathbf{H})\} + \frac{\theta T B n_R}{\log_e 2}\left(\mathbb{E}\{\lambda^2_{\max}(\mathbf{H}^\dagger\mathbf{H})\} - \mathbb{E}^2\{\lambda_{\max}(\mathbf{H}^\dagger\mathbf{H})\}\right)} && (58) \\ &= \frac{2}{\frac{n_R}{l}\kappa(\sigma_{\max}(\mathbf{H})) + \frac{\theta T B n_R}{\log_e 2}\left(\kappa(\sigma_{\max}(\mathbf{H})) - 1\right)} && (59) \end{aligned}$$

where $\kappa(\sigma_{\max}(H))$ is the kurtosis of maximum singular value of the matrix $\mathbf{H}$ and is defined as

$$\kappa(\sigma_{\max}(\mathbf{H})) = \frac{\mathbb{E}\{\sigma^4_{\max}(\mathbf{H})\}}{\mathbb{E}^2\{\sigma^2_{\max}(\mathbf{H})\}} = \frac{\mathbb{E}\{\lambda^2_{\max}(\mathbf{H}^\dagger\mathbf{H})\}}{\mathbb{E}^2\{\lambda_{\max}(\mathbf{H}^\dagger\mathbf{H})\}}. \tag{60}$$

Remark 2: In [2], Shannon capacity is considered and it is shown that $\frac{E_b}{N_0}_{\min} = \frac{\log_e 2}{\mathbb{E}\{\lambda_{\max}(\mathbf{H}^\dagger\mathbf{H})\}}$ and $\mathcal{S}_0 = \frac{2}{\frac{n_R}{l}\kappa(\sigma_{\max}(\mathbf{H}))}$. From (57) and (59) we note that we have the same minimum bit energy in the presence of QoS limitations while the wideband slope diminishes with increasing θ.

When we have equal power allocation, i.e., $\mathbf{K}_x = \frac{1}{n_T}\mathbf{I}$, it can be immediately seen from the result of Corollary 1 that

$$\frac{E_b}{N_0}_{\min} = \frac{n_T \log_e 2}{\mathbb{E}\{\mathrm{tr}\,(\mathbf{H}^\dagger\mathbf{H})\}} \tag{61}$$

$$\mathcal{S}_0 = \frac{2\mathbb{E}^2\{\mathrm{tr}\,(\mathbf{H}^\dagger\mathbf{H})\}}{n_R\mathbb{E}\{\mathrm{tr}\,((\mathbf{H}^\dagger\mathbf{H})^2)\} + \frac{\theta T B n_R}{\log_e 2}\left(\mathbb{E}\{\mathrm{tr}^2(\mathbf{H}^\dagger\mathbf{H})\} - \mathbb{E}^2\{\mathrm{tr}\,(\mathbf{H}^\dagger\mathbf{H})\}\right)}. \tag{62}$$

[5] We note that the expressions in (54) and (56) differ from those in [2] by a constant factor due to our assumption that the units of C_E is bits/s/Hz/dimension rather than nats/s/Hz/dimension.

Assume that $\mathbf{H}$ has independent zero-mean unit-variance complex Gaussian random entries. Under this assumption, we have [3]

$$\mathbb{E}\{\text{tr}\,(\mathbf{H}^\dagger\mathbf{H})\} = n_R n_T, \quad \mathbb{E}\{\text{tr}^{\,2}(\mathbf{H}^\dagger\mathbf{H})\} = n_R n_T(n_R n_T + 1), \quad \mathbb{E}\{\text{tr}\,((\mathbf{H}^\dagger\mathbf{H})^2)\} = n_R n_T(n_R + n_T). \tag{63}$$

Using these facts, we have the following minimum bit energy and wideband slope expressions for the uniform power allocation case when the entries of $\mathbf{H}$ are i.i.d. zero-mean unit-variance Gaussian random variables:

$$\frac{E_b}{N_0}_{\text{min}} = \frac{\log_e 2}{n_R} \quad \text{and} \quad \mathcal{S}_0 = \frac{2}{\frac{n_R+n_T}{n_T} + \frac{\theta TB}{n_T \log_e 2}} \quad \text{bits/s/Hz/(3 dB)/receive antenna}. \tag{64}$$

We note that while the minimum bit energy depends only on the number of receive antennas, the wideband slope is a function of both the receive and transmit antennas. Note that the wideband slope expression is per receive antenna. Without this normalization, we have

$$\mathcal{S}_0 = \frac{2}{\frac{n_R+n_T}{n_r n_T} + \frac{\theta TB}{n_R n_T \log_e 2}} \quad \text{bits/s/Hz/(3 dB)}. \tag{65}$$

From (65), we identify the interactions between the spatial dimensions and QoS constraints. Note that more strict QoS constraints and hence higher values of θ tend to diminish the wideband slope. On the other hand, we see in the second term in the denominator of (65) that the impact of the presence of QoS constraints is being diminished by the product of the number of transmit and receive antennas, $n_R n_T$. Hence, increasing the number of transmit and/or receive antennas can offset the performance loss due to queueing constraints.

V. Minimum Bit Energy in the Wideband Regime

In the previous section, we have assumed that the bandwidth of the system is fixed as the transmission power P diminishes and system operates in the low-power regime. Here, we study the regime in which the bandwidth increases while P is kept fixed. Note that as the bandwidth grows, the flat-fading assumption will no longer hold and the input-output relation given in (1) will not be an accurate description. On the other hand, if we decompose the wideband channel into parallel, noninteracting, narrowband subchannels each with bandwidth that is equal to the coherence bandwidth B_c, then we can assume that each subchannel experiences independent flat fading and has an input-output relation that can be expressed as

$$\mathbf{y}_i = \mathbf{H}_i \mathbf{x}_i + \mathbf{n}_i \quad i = 1, 2, \ldots, m \tag{66}$$

where $\mathbf{x}_i$ and $\mathbf{y}_i$ are the input and output vectors of the i^{th} subchannel, and $\mathbf{H}_i$ is the i^{th} subchannel matrix. $\mathbf{n}_i$ represents the additive zero mean Gaussian noise vector with $E\{\mathbf{n}_i \mathbf{n}_i^\dagger\} = N_0 \mathbf{I}$ in the i^{th} subchannel. We assume that the input in the i^{th} subchannel is subject to $E\{\|\mathbf{x}_i\|^2\} \le \frac{P_i}{B_c}$ where P_i is the power allocated to the i^{th} subchannel. We assume that the number of subchannels is m and hence we have $B = mB_c$ and $\sum_{i=1}^m P_i = P$ where B and P denote the total bandwidth and power, respectively, of the wideband system. Under these assumptions, the maximum instantaneous transmission rate in the i^{th} subchannel with covariance matrix $\mathbf{K}_{x,i}$ is

$$B_c \log_2 \det \left(\mathbf{I} + n_R \text{SNR}_i \mathbf{H}_i \mathbf{K}_{x,i} \mathbf{H}_i^\dagger \right) \text{ bits/s} \tag{67}$$

where $\text{SNR}_i = \frac{P_i}{n_R B_c N_0}$. Due to the independence of fading in different subchannels, the total transmission rate over the wideband channel is

$$\sum_{i=1}^{m} B_c \log_2 \det \left(\mathbf{I} + n_R \text{SNR}_i \mathbf{H}_i \mathbf{K}_{x,i} \mathbf{H}_i^\dagger \right) \text{ bits/s} \tag{68}$$

which is achieved by independent signaling over different subchannels, i.e., by choosing $\{\mathbf{x}_i\}_{i=1}^m$ as zero-mean independent Gaussian vectors with covariance matrices $\{\mathbf{K}_{x,i}\}_{i=1}^m$. Then, for the transmission rate in (68), the effective rate is given by

$$\begin{aligned}
\mathsf{R}_E(\text{SNR}) &= -\frac{1}{\theta T B n_R} \log_e \mathbb{E} \left\{ \exp \left(-\theta T B_c \sum_{i=1}^{m} \log_2 \det \left(\mathbf{I} + n_R \text{SNR}_i \mathbf{H}_i \mathbf{K}_{x,i} \mathbf{H}_i^\dagger \right) \right) \right\} && (69) \\
&= -\frac{1}{\theta T B n_R} \log_e \prod_{i=1}^{m} \mathbb{E} \left\{ \exp \left(-\theta T B_c \log_2 \det \left(\mathbf{I} + n_R \text{SNR}_i \mathbf{H}_i \mathbf{K}_{x,i} \mathbf{H}_i^\dagger \right) \right) \right\} && (70) \\
&= -\frac{1}{\theta T B n_R} \sum_{i=1}^{m} \log_e \mathbb{E} \left\{ \exp \left(-\theta T B_c \log_2 \det \left(\mathbf{I} + n_R \text{SNR}_i \mathbf{H}_i \mathbf{K}_{x,i} \mathbf{H}_i^\dagger \right) \right) \right\} && (71)
\end{aligned}$$

where (70) follows from our assumption that $\{\mathbf{H}_i\}$ are independent subchannel matrices and the fact that the expected value of a product of independent random variables is equal to the product of the expected values of the individual random variables. In general, effective capacity can be obtained by maximizing the effective rate expression in (71) over all power allocations $\{P_i\}$ and covariance matrices $\{\mathbf{K}_{x,i}\}$. If the channel is known at the transmitter, $\{P_i\}$ and $\{\mathbf{K}_{x,i}\}$ can depend on the realizations of the channel matrices $\{\mathbf{H}_i\}$.

We simplify the above setting by assuming that $\mathbf{H}_i\mathbf{K}_{x,i}\mathbf{H}_i^\dagger$ has the same distribution for all $i = 1, 2, \ldots, m$. For instance, this assumption would hold when $\{\mathbf{H}_i\}$ are identically distributed, and $\mathbf{K}_{x,i}$ is the same fixed matrix for all i or is a random matrix with a common distribution for all i (e.g., $\mathbf{K}_{x,i} = \mathbf{u}\mathbf{u}^\dagger$, where $\mathbf{u}$ is the random eigenvector that corresponds to $\lambda_{\max}(\mathbf{H}_i^\dagger\mathbf{H}_i)$, has the same distribution for all i when $\{\mathbf{H}_i\}$ are identically distributed). Under this assumption, we can eliminate the dependence of $\mathbf{H}_i\mathbf{K}_{x,i}\mathbf{H}_i^\dagger$ on the time index i, and show from the concavity of the expression (71) with respect to signal-to-noise ratio[6] that the effective rate is maximized by having $\text{SNR}_i = \frac{P/m}{n_R N_0 B_c} = \frac{P}{n_R N_0 B} = \text{SNR}$ for all i, i.e., by distributing the total power equally over the subchannels. Now, the effective rate expression becomes

$$\mathsf{R}_E(\text{SNR}) = -\frac{1}{\theta T B n_R} m \log_e \mathbb{E}\left\{\exp\left(-\theta T B_c \log_2 \det\left(\mathbf{I} + n_R \text{SNR}\mathbf{H}\mathbf{K}_x\mathbf{H}^\dagger\right)\right)\right\} \tag{72}$$

$$= -\frac{1}{\theta T B_c n_R} \log_e \mathbb{E}\left\{\exp\left(-\theta T B_c \log_2 \det\left(\mathbf{I} + n_R \text{SNR}\mathbf{H}\mathbf{K}_x\mathbf{H}^\dagger\right)\right)\right\} \tag{73}$$

where we have used the relation $B = mB_c$.

Now, we analyze the effective capacity and energy efficiency in the wideband limit in three scenarios:

1) Rich Multipath Fading: In a system with bandwidth B, the maximum number of resolvable paths is proportional to $BT_m = \frac{B}{B_c}$ where T_m denotes the delay spread and $B_c = \frac{1}{T_m}$. In rich multipath fading, the assumption is that the number of independent resolvable paths increases linearly with increasing bandwidth. Therefore, in rich multipath fading, coherence bandwidth B_c remains fixed as B increases while $\text{SNR} = \frac{P}{BN_0}$ diminishes to zero. Then, from the similarity of the effective rate expressions in (16) and (73) and the fact that B is fixed in (16) in the low-power regime analysis, we immediately conclude that the wideband and low-power results are identical in rich multipath fading under the assumptions that lead to the effective rate expression in (73).

2) Sparse Multipath Fading: In sparse multipath fading, it is assumed that the number of independent resolvable paths increases at most *sublinearly* with bandwidth [23] [24]. Hence, in this case, B_c increases with increasing bandwidth. In the special case in which the number of resolvable paths is bounded, B_c increases linearly with B while the number of subchannels m remains fixed. For instance, such a scenario is considered in [25]. For this case, we have the following result on the minimum bit energy required in the wideband regime.

[6] Since $-\theta T B_c \log_2 \det\left(\mathbf{I} + n_R \text{SNR}_i \mathbf{H}_i\mathbf{K}_{x,i}\mathbf{H}_i^\dagger\right)$ is a convex function of SNR for given $\mathbf{H}_i\mathbf{K}_{x,i}\mathbf{H}_i^\dagger$, $e^{-\theta T B_c \log_2 \det\left(\mathbf{I} + n_R \text{SNR}_i \mathbf{H}_i\mathbf{K}_{x,i}\mathbf{H}_i^\dagger\right)}$ is a log-convex function. Moreover, since log-convexity is preserved under sums [32, Section 3.5.2], $\mathbb{E}\left\{\exp\left(-\theta T B_c \log_2 \det\left(\mathbf{I} + n_R \text{SNR}_i \mathbf{H}_i\mathbf{K}_{x,i}\mathbf{H}_i^\dagger\right)\right)\right\}$ is log-convex, implying that $\log_e \mathbb{E}\left\{\exp\left(-\theta T B_c \log_2 \det\left(\mathbf{I} + n_R \text{SNR}_i \mathbf{H}_i\mathbf{K}_{x,i}\mathbf{H}_i^\dagger\right)\right)\right\}$ is a convex function of SNR. Since the sum of convex functions is convex [32], and the negative of a convex function is concave, we conclude that the expression in (71) is a concave function of SNR.

Theorem 3: Assume that the number of independent resolvable paths remain bounded and fixed in the wideband regime as B increases. In this case, the minimum bit energy for a given covariance matrix $\mathbf{K}_x$ is given by

$$\frac{E_b}{N_0}_{\min} = \frac{\frac{\theta T P}{m N_0}}{-\log_e \mathbb{E}\left\{e^{-\frac{\theta T P}{m N_0}\frac{1}{\log_e 2}\,\text{tr}\,(\mathbf{H}\mathbf{K}_x\mathbf{H}^\dagger)}\right\}}. \tag{74}$$

When the channel is perfectly known at the transmitter, information can be sent in the maximal-eigenvalue eigenspace of $\mathbf{H}^\dagger\mathbf{H}$ and the required minimum bit energy becomes

$$\frac{E_b}{N_0}_{\min} = \frac{\frac{\theta T P}{m N_0}}{-\log_e \mathbb{E}\left\{e^{-\frac{\theta T P}{m N_0}\frac{1}{\log_e 2}\lambda_{\max}(\mathbf{H}^\dagger\mathbf{H})}\right\}}. \tag{75}$$

If only statistical information of the channel is available at the transmitter, the minimum bit energy can be obtained by minimizing (74) over all permissible covariance matrices, i.e.,

$$\frac{E_b}{N_0}_{\min} = \min_{\substack{\mathbf{K}_x \succeq \mathbf{0} \\ \text{tr}\,(\mathbf{K}_x)\le 1}} \frac{\frac{\theta T P}{m N_0}}{-\log_e \mathbb{E}\left\{e^{-\frac{\theta T P}{m N_0}\frac{1}{\log_e 2}\,\text{tr}\,(\mathbf{H}\mathbf{K}_x\mathbf{H}^\dagger)}\right\}}. \tag{76}$$

Proof: For a given input covariance matrix $\mathbf{K}_x$, the bit energy required for reliable communications under QoS constraints is

$$\frac{E_b}{N_0} = \frac{\text{SNR}}{\mathsf{R}_E(\text{SNR})} = \frac{\frac{P}{n_R B N_0}}{-\frac{1}{\theta T B_c n_R}\log_e \mathbb{E}\left\{\exp\left(-\theta T B_c \log_2 \det\left(\mathbf{I} + n_R \text{SNR}\mathbf{H}\mathbf{K}_x\mathbf{H}^\dagger\right)\right)\right\}} \tag{77}$$

$$= \frac{\frac{\theta T P}{m N_0}}{-\log_e \mathbb{E}\left\{\exp\left(-\theta T B_c \log_2 \det\left(\mathbf{I} + \frac{P}{m B_c N_0}\mathbf{H}\mathbf{K}_x\mathbf{H}^\dagger\right)\right)\right\}} \tag{78}$$

$$= \frac{\frac{\theta T P}{m N_0}}{-\log_e \mathbb{E}\left\{\exp\left(-\theta T B_c \sum_i \log_2\left(1 + \frac{P}{m B_c N_0}\lambda_i(\mathbf{H}\mathbf{K}_x\mathbf{H}^\dagger)\right)\right)\right\}} \tag{79}$$

where $\lambda_i(\mathbf{H}\mathbf{K}_x\mathbf{H}^\dagger)$ denotes the i^{th} eigenvalue of the matrix $\mathbf{H}\mathbf{K}_x\mathbf{H}^\dagger$. Above, (78) is obtained by using the relation $B = mB_c$ and performing some straightforward algebraic operations, and (79) follows from the fact that $\det(\mathbf{A}) = \prod_i \lambda_i(\mathbf{A})$. Note that under the assumption of fixed number of resolvable paths, B_c increases linearly with B while m is fixed. Hence, only the denominator of (79) varies with B. From the fact that the function $x \log_2(1 + \frac{a}{x})$ is a monotonically increasing function of $x > 0$ for any constant $a > 0$, we can easily see that the minimum bit energy is achieved as $B \to \infty$. Since B_c also grows without bound as B increases, we have

$$\frac{E_b}{N_0}_{\min} = \lim_{B_c \to \infty} \frac{\frac{\theta T P}{m N_0}}{-\log_e \mathbb{E}\left\{\exp\left(-\theta T B_c \sum_i \log_2\left(1 + \frac{P}{m B_c N_0} \lambda_i(\mathbf{H}\mathbf{K}_x\mathbf{H}^\dagger)\right)\right)\right\}} \tag{80}$$

$$= \frac{\frac{\theta T P}{m N_0}}{-\log_e \mathbb{E}\left\{\exp\left(-\theta T \frac{1}{\log_e 2} \sum_i \frac{P}{m N_0} \lambda_i(\mathbf{H}\mathbf{K}_x\mathbf{H}^\dagger)\right)\right\}} \tag{81}$$

$$= \frac{\frac{\theta T P}{m N_0}}{-\log_e \mathbb{E}\left\{\exp\left(-\frac{\theta T P}{m N_0} \frac{1}{\log_e 2} \sum_i \lambda_i(\mathbf{H}\mathbf{K}_x\mathbf{H}^\dagger)\right)\right\}} \tag{82}$$

$$= \frac{\frac{\theta T P}{m N_0}}{-\log_e \mathbb{E}\left\{\exp\left(-\frac{\theta T P}{m N_0} \frac{1}{\log_e 2} \operatorname{tr}\left(\mathbf{H}\mathbf{K}_x\mathbf{H}^\dagger\right)\right)\right\}}. \tag{83}$$

(81) is obtained using the fact that as $B_c \to \infty$, we have $B_c \log_2\left(1 + \frac{P}{m B_c N_0} \lambda_i(\mathbf{H}\mathbf{K}_x\mathbf{H}^\dagger)\right) \to \frac{1}{\log_e 2} \frac{P}{m N_0} \lambda_i(\mathbf{H}\mathbf{K}_x\mathbf{H}^\dagger)$. (83) follows from the property that $\sum_i \lambda_i(\mathbf{A}) = \operatorname{tr}(\mathbf{A})$. Note that (83) proves (74) which is the minimum bit energy for a given covariance matrix $\mathbf{K}_x$.

Recall that it is shown in the proof of Theorem 1 that

$$\operatorname{tr}\left(\mathbf{H}\mathbf{K}_x\mathbf{H}^\dagger\right) \le \lambda_{\max}(\mathbf{H}^\dagger\mathbf{H}) \tag{84}$$

and this upper bound can be achieved by transmitting in the maximal-eigenvalue eigenspace of $\mathbf{H}\mathbf{H}^\dagger$, e.g., by having $\mathbf{K}_x = \mathbf{u}\mathbf{u}^\dagger$ where $\mathbf{u}$ is the eigenvector that corresponds to $\lambda_{\max}(\mathbf{H}^\dagger\mathbf{H})$. If the transmitter perfectly knows the realizations of the channel matrix $\mathbf{H}$, then this transmission strategy can be employed and the minimum bit energy becomes

$$\frac{E_b}{N_0}_{\min} = \frac{\frac{\theta T P}{m N_0}}{-\log_e \mathbb{E}\left\{\exp\left(-\frac{\theta T P}{m N_0} \frac{1}{\log_e 2} \lambda_{\max}(\mathbf{H}\mathbf{H}^\dagger)\right)\right\}}. \tag{85}$$

If the transmitter has only statistical knowledge of the channel matrix, the minimum bit energy can be determined by finding the input covariance matrix that minimizes (83). ■

Remark 3: By applying the Jensen's inequality, we can easily see that

$$\log_e \mathbb{E}\left\{e^{-\frac{\theta T P}{m N_0} \frac{1}{\log_e 2} \operatorname{tr}\left(\mathbf{H}\mathbf{K}_x\mathbf{H}^\dagger\right)}\right\} \ge \mathbb{E}\left\{\log_e e^{-\frac{\theta T P}{m N_0} \frac{1}{\log_e 2} \operatorname{tr}\left(\mathbf{H}\mathbf{K}_x\mathbf{H}^\dagger\right)}\right\} = \mathbb{E}\left\{-\frac{\theta T P}{m N_0} \frac{1}{\log_e 2} \operatorname{tr}\left(\mathbf{H}\mathbf{K}_x\mathbf{H}^\dagger\right)\right\} \tag{86}$$

which implies that

$$\frac{E_b}{N_0}_{\min} = \frac{\frac{\theta T P}{m N_0}}{-\log_e \mathbb{E}\left\{e^{-\frac{\theta T P}{m N_0} \frac{1}{\log_e 2} \operatorname{tr}\left(\mathbf{H}\mathbf{K}_x\mathbf{H}^\dagger\right)}\right\}} \ge \frac{\log_e 2}{\operatorname{tr}\left(\mathbf{H}\mathbf{K}_x\mathbf{H}^\dagger\right)}. \tag{87}$$

Similarly, we can show

$$\frac{E_b}{N_0}_{\min} = \frac{\frac{\theta T P}{m N_0}}{-\log_e \mathbb{E}\left\{e^{-\frac{\theta T P}{m N_0}\frac{1}{\log_e 2}\lambda_{\max}(\mathbf{H}^\dagger\mathbf{H})}\right\}} \geq \frac{\log_e 2}{\lambda_{\max}(\mathbf{H}^\dagger\mathbf{H})}. \tag{88}$$

$$\frac{E_b}{N_0}_{\min} = \min_{\substack{\mathbf{K}_x\succeq\mathbf{0}\\ \text{tr}\,(\mathbf{K}_x)\leq 1}} \frac{\frac{\theta T P}{m N_0}}{-\log_e \mathbb{E}\left\{e^{-\frac{\theta T P}{m N_0}\frac{1}{\log_e 2}\text{tr}\,(\mathbf{H}\mathbf{K}_x\mathbf{H}^\dagger)}\right\}} \geq \min_{\substack{\mathbf{K}_x\succeq\mathbf{0}\\ \text{tr}\,(\mathbf{K}_x)\leq 1}} \frac{\log_e 2}{\text{tr}\,(\mathbf{H}\mathbf{K}_x\mathbf{H}^\dagger)} = \frac{\log_e 2}{\lambda_{\max}(\mathbb{E}\{\mathbf{H}^\dagger\mathbf{H}\})} \tag{89}$$

Note that the right-hand sides of the above inequalities are the minimum bit energy expressions in the low-power regime and also the wideband regime with rich multipath fading due to the equivalence of the two. From this, we immediately conclude that the sparse multipath fading with bounded number of resolvable paths (or equivalently bounded number of subchannels) induces additional energy requirements in the presence of QoS constraints.

Remark 4: Recall from the result of Theorem 2 that when the transmitter has only statistical knowledge of the channel, the optimal transmission strategy in the low-power regime (and also in the wideband regime with rich multipath fading) is to transmit the information in the maximal-eigenvalue eigenspace of $\mathbb{E}\{\mathbf{H}^\dagger\mathbf{H}\}$. On the other hand, we note from Theorem 3 that this is not necessarily the optimal transmission technique in the wideband regime with sparse fading. The optimal input covariance is the one that minimizes (74). Note further that for small $\frac{\theta T P}{m N_0}$, we have the following first-order Taylor series expansion of the denominator of (74):

$$-\log_e \mathbb{E}\left\{e^{-\frac{\theta T P}{m N_0}\frac{1}{\log_e 2}\text{tr}\,(\mathbf{H}\mathbf{K}_x\mathbf{H}^\dagger)}\right\} = \frac{\theta T P}{m N_0}\frac{1}{\log_e 2}\text{tr}\,(\mathbf{H}\mathbf{K}_x\mathbf{H}^\dagger) + o\left(\frac{\theta T P}{m N_0}\right). \tag{90}$$

Hence, when θ or P is small or m is large, the input covariance that is optimal to the first order is the one that maximizes $\text{tr}\,(\mathbf{H}\mathbf{K}_x\mathbf{H}^\dagger)$, i.e., in this case, transmission in the maximal-eigenvalue eigenspace of $\mathbb{E}\{\mathbf{H}^\dagger\mathbf{H}\}$ is optimal as in the low-power regime.

Theorem 3 holds for the case in which the number of resolvable multipath components remains bounded. Another scenario in sparse multipath fading is the one in which the number of resolvable paths increases with bandwidth but only *sublinearly*. In this case, both B_c and m increase without bound as $B \to \infty$ due to the sublinear growth of B_c. Therefore, the minimum bit energy results can be obtained by letting $m \to \infty$ in the results of Theorem 3.

Theorem 4: Assume a sparse multipath fading scenario in which the number of independent resolvable paths increase sublinearly with bandwidth. In this case, the minimum bit energy for a given input covariance matrix is given by

$$\frac{E_b}{N_0}_{\min} = \frac{\log_e 2}{\text{tr}\,(\mathbf{H}\mathbf{K}_x\mathbf{H}^\dagger)}. \tag{91}$$

When the transmitter perfectly knows the channel matrix $\mathbf{H}$ and when it knows only $\mathbb{E}\{\mathbf{H}^\dagger\mathbf{H}\}$, the minimum bit energies are

$$\frac{E_b}{N_0}_{\min} = \frac{\log_e 2}{\lambda_{\max}(\mathbf{H}^\dagger\mathbf{H})} \quad \text{and} \quad \frac{E_b}{N_0}_{\min} = \frac{\log_e 2}{\lambda_{\max}(\mathbb{E}\{\mathbf{H}^\dagger\mathbf{H}\})}, \tag{92}$$

respectively.

Proof: As mentioned above, proof follows by finding the limiting values of the minimum bit energy expressions in Theorem 3 as $m \to \infty$. For the case of fixed covariance matrix $\mathbf{K}_x$, we have

$$\begin{aligned}
\frac{E_b}{N_0}_{\min} &= \lim_{m\to\infty} \frac{\frac{\theta TP}{mN_0}}{-\log_e \mathbb{E}\left\{e^{-\frac{\theta TP}{mN_0}\frac{1}{\log_e 2}\text{tr}\,(\mathbf{H}\mathbf{K}_x\mathbf{H}^\dagger)}\right\}} &(93)\\
&= \lim_{m\to\infty} \frac{\frac{\theta TP}{mN_0}}{\frac{\theta TP}{mN_0}\frac{1}{\log_e 2}\text{tr}\,(\mathbf{H}\mathbf{K}_x\mathbf{H}^\dagger) + o\left(\frac{\theta TP}{mN_0}\right)} &(94)\\
&= \lim_{m\to\infty} \frac{1}{\frac{1}{\log_e 2}\text{tr}\,(\mathbf{H}\mathbf{K}_x\mathbf{H}^\dagger) + \frac{o\left(\frac{\theta TP}{mN_0}\right)}{\frac{\theta TP}{mN_0}}} &(95)\\
&= \frac{1}{\frac{1}{\log_e 2}\text{tr}\,(\mathbf{H}\mathbf{K}_x\mathbf{H}^\dagger) + \lim_{m\to\infty}\frac{o\left(\frac{\theta TP}{mN_0}\right)}{\frac{\theta TP}{mN_0}}} &(96)\\
&= \frac{\log_e 2}{\text{tr}\,(\mathbf{H}\mathbf{K}_x\mathbf{H}^\dagger)}. &(97)
\end{aligned}$$

(94) is obtained by using the first-order Taylor expansion in (90). (95) follows by dividing the numerator and denominator by $\frac{\theta TP}{mN_0}$. Finally, (97) is obtained immediately from the definition that $\lim_{x\to 0}\frac{o(x)}{x} = 0$. The expressions in (92) are determined as in the proofs of Theorems 1 and 2 by choosing the input covariance matrix as $\mathbf{K}_x = \mathbf{u}\mathbf{u}^\dagger$ where $\mathbf{u}$ is the eigenvector that corresponds to the maximum eigenvalue of $\mathbf{H}^\dagger\mathbf{H}$ (when $\mathbf{H}$ is perfectly known at the transmitter) or of $\mathbb{E}\{\mathbf{H}^\dagger\mathbf{H}\}$ (when only $\mathbb{E}\{\mathbf{H}^\dagger\mathbf{H}\}$ is known at the transmitter). ■

Remark 5: Theorem 4 shows that as long as the number of subchannels m grows without bound in the wideband regime, the minimum bit energy requirements are the same as those in the low-power regime and wideband regime with rich multipath fading in which m increases linearly with bandwidth. Note that since each subchannel experiences independent fading, m can be seen as a measure of the degrees of freedom in the system. Therefore, if m is bounded, the degrees of freedom is also bounded and that results in increased energy requirements as discussed in Remark 3. On the other hand, if the degrees of freedom increase with bandwidth, we have the same minimum bit energy values even though the increase is sublinear. However, for this case, we will observe in the numerical results in Section VII that approaching the minimum bit energy is very slow and demanding in bandwidth due to zero wideband slope.

Remark 6: Note that having $m \to \infty$ for fixed $\theta > 0$ in the minimum bit energy expressions in (74)–(76) is the same as letting $\theta \to 0$ for fixed m. Hence, even if m is bounded, the minimum bit energies given in Theorem 4 are attained when $\theta = 0$. This indicates that multipath sparsity does not affect the performance in the absence of QoS constraints.

VI. The Impact of QoS Constraints in the High-SNR Regime

In this section, we consider a single flat-fading channel and analyze how QoS limitations affect the performance in the high-SNR regime. In contrast to the previous sections where general models are used, we here consider a specific fading scenario in which the components of $\mathbf{H}$ are independent and identically distributed (i.i.d.) Gaussian random variables with zero mean and unit variance. Moreover, we assume that the input covariance matrix is $\mathbf{K}_x = \frac{1}{n_T}\mathbf{I}$. Note that this covariance matrix is optimal in the sense of achieving the ergodic Shannon capacity when $\mathbf{H}$ has the above distribution and the transmitter does not know the realizations of $\mathbf{H}$ [26].

Now, for the considered channel and input models, the effective rate is given by

$$R_{E,\text{id}}(\text{SNR}) = -\frac{1}{\theta TB}\log_e \mathbb{E}\left\{\exp\left(-\theta TB \log_2 \det\left(\mathbf{I} + \frac{n_R}{n_T}\text{SNR}\mathbf{H}\mathbf{H}^\dagger\right)\right)\right\} \text{ bits/s/Hz.} \tag{98}$$

Note that in the above formulation, we have not normalized the effective rate expression with the number of receive antennas n_R, and we have chosen a slightly different font from before and use the notation $R_{E,\text{id}}$ to denote this unnormalized effective rate.

As also pointed before, the effective capacity and effective rate expressions are proportional to the logarithm of the moment generating functions of instantaneous transmission rates. For the channel and input models considered in this section, Wang and Giannakis in [28, Theorem 1] provided an expression for the moment generating function of instantaneous mutual information. Applying this result to our setting, we obtain

$$\mathbb{E}\left\{\exp\left(-\theta TB \log_2 \det\left(\mathbf{I} + \frac{n_R}{n_T}\text{SNR}\mathbf{H}\mathbf{H}^\dagger\right)\right)\right\} = \frac{\det(\mathbf{G}(\theta, \text{SNR}))}{\prod_{i=1}^{k}\Gamma(d+i)} \tag{99}$$

where $\Gamma(\cdot)$ is the Gamma function, $k = \min(n_R, n_T)$, and $d = \max(n_R, n_T) - \min(n_R, n_T)$. Moreover,

$\mathbf{G}$ is a $k \times k$ Hankel matrix whose $(i,j)^{\text{th}}$ component is

$$g_{i,j} = \int_0^\infty \left(1 + \frac{n_R}{n_T}\,\text{SNR}\,z\right)^{-\theta TB \log_2 e} z^{i+j+d}\, e^{-z}\, dz \qquad i,j = 0,1,\ldots,k-1. \tag{100}$$

Therefore, we have

$$R_{E,\text{id}}(\text{SNR}) = -\frac{1}{\theta TB} \log_e \left(\frac{\det(\mathbf{G}(\theta, \text{SNR}))}{\prod_{i=1}^k \Gamma(d+i)} \right). \tag{101}$$

In order to quantify the impact of the QoS constraints on the performance in the high-SNR regime, we consider two measures, $\mathcal{S}_\infty$ and $\mathcal{L}_\infty$, which are defined as

$$\mathcal{S}_\infty = \lim_{\text{SNR}\to\infty} \frac{R_{E,\text{id}}(\text{SNR})}{\log_2 \text{SNR}} \tag{102}$$

and

$$\mathcal{L}_\infty = \lim_{\text{SNR}\to\infty} \left(\log_2 \text{SNR} - \frac{R_{E,\text{id}}(\text{SNR})}{\mathcal{S}_\infty} \right). \tag{103}$$

Note that while $\mathcal{S}_\infty$ denotes the high-SNR slope in bits/s/Hz/(3dB), $\mathcal{L}_\infty$ represents the power offset with respect to a reference channel having the same high-SNR slope but with unfaded and orthogonal dimensions [4]. With these quantities, the effective rate is approximated at high SNRS as

$$R_{E,\text{id}} = \mathcal{S}_\infty(\log_2 \text{SNR} - \mathcal{L}_\infty) + o(1). \tag{104}$$

The above high-SNR approximation was first introduced and used in [27] in the study of code-division multiple access systems with random spreading, and was later employed in [4] in the study of ergodic Shannon capacity of multiple-antenna systems. Here, we apply this approximation to the multiple-antenna systems operating under statistical queueing constraints. The next result identifies the values of $\mathcal{S}_\infty$ and $\mathcal{L}_\infty$ for a subset of values of the QoS exponent θ.

Theorem 5: Assume that the components of channel matrix $\mathbf{H}$ are independent and identically distributed (i.i.d.) Gaussian random variables with zero mean and unit variance. If the QoS exponent satisfies

$$\theta < \frac{\max(n_R, n_T) - \min(n_R, n_T) + 1}{TB \log_2 e}, \tag{105}$$

then, we have

$$\mathcal{S}_\infty = \min(n_R, n_T), \tag{106}$$

and

$$\mathcal{L}_\infty = \begin{cases} \log_2 \frac{n_T}{n_R} + \frac{1}{\theta T B n_R} \log_e E\left\{ e^{-\theta T B \log_2 \det \mathbf{H}\mathbf{H}^\dagger} \right\} & n_R \le n_T \\ \log_2 \frac{n_T}{n_R} + \frac{1}{\theta T B n_T} \log_e E\left\{ e^{-\theta T B \log_2 \det \mathbf{H}^\dagger\mathbf{H}} \right\} & n_R > n_T \end{cases}. \tag{107}$$

Proof: Note that the components of the Hankel matrix $\mathbf{G}$, which appears in the effective rate expression in (101), can be written as

$$g_{i,j} = \text{SNR}^{-\theta T B \log_2 e} \int_0^\infty \left(\frac{1}{\text{SNR}} + \frac{n_R}{n_T} z \right)^{-\theta T B \log_2 e} z^{i+j+d} \, e^{-z} \, dz \qquad i,j = 0,1,\ldots,k-1 \tag{108}$$

where $k = \min(n_R, n_T)$. As $\text{SNR} \to \infty$, the integral in the above expression goes to a nonzero and finite value if $-\theta T B \log_2 e + i + j + d > -1$ since $0 < \int_0^\infty z^a e^{-z} dz < \infty$ for $a > -1$ and $\int_0^\infty z^a e^{-z} dz = \infty$ for $a \le -1$. Note that this condition is satisfied for all $i, j = 0, 1, \ldots, k-1$ by our assumption in (105). Now, we can immediately see that $g_{i,j}$ for all i, j scales as $\text{SNR}^{-\theta T B \log_2 e}$ as $\text{SNR} \to \infty$. Therefore, the determinant of $\mathbf{G}$ scales as $\text{SNR}^{-k\theta T B \log_2 e}$. This lets us conclude that

$$\begin{aligned} R_{E,\text{id}}(\text{SNR}) = -\frac{1}{\theta T B} \log_e \left(\frac{\det(\mathbf{G}(\theta, \text{SNR}))}{\prod_{i=1}^k \Gamma(d+i)} \right) &= -\frac{1}{\theta T B} \log_e \left(\text{SNR}^{-k\theta T B \log_2 e} \right) + O(1) & (109) \\ &= k(\log_2 e) \log_e \text{SNR} + O(1) & (110) \\ &= k \log_2 \text{SNR} + O(1) & (111) \\ &= \min(n_R, n_T) \log_2 \text{SNR} + O(1), & (112) \end{aligned}$$

establishing that $\mathcal{S}_\infty = \min(n_R, n_T)$ for the values of θ specified in the theorem. Above, $O(1)$ denotes the terms that approach a finite constant as $\text{SNR} \to \infty$.

Next, we consider the power offset $\mathcal{L}_\infty$. Assume that $n_R \le n_T$. Under this assumption, we have

$$\begin{aligned} \mathcal{L}_\infty &= \lim_{\text{SNR}\to\infty} \left(\log_2 \text{SNR} - \frac{R_{E,\text{id}}(\text{SNR})}{\mathcal{S}_\infty} \right) & (113) \\ &= \lim_{\text{SNR}\to\infty} \left(\log_2 \text{SNR} - \frac{R_{E,\text{id}}(\text{SNR})}{n_R} \right) & (114) \\ &= \lim_{\text{SNR}\to\infty} \left(\log_2 \text{SNR} + \frac{\frac{1}{\theta T B} \log_e \mathbb{E}\left\{ e^{-\theta T B \log_2 \det\left(\mathbf{I} + \frac{n_R}{n_T} \text{SNR}\, \mathbf{H}\mathbf{H}^\dagger\right)} \right\}}{n_R} \right) & (115) \\ &= \lim_{\text{SNR}\to\infty} \left(\log_2 \text{SNR} + \frac{\frac{1}{\theta T B} \log_e \mathbb{E}\left\{ e^{-\theta T B n_R \log_2 \text{SNR} - \theta T B \log_2 \det\left(\frac{1}{\text{SNR}}\mathbf{I} + \frac{n_R}{n_T} \mathbf{H}\mathbf{H}^\dagger\right)} \right\}}{n_R} \right) & (116) \end{aligned}$$

$$= \lim_{\text{SNR}\to\infty} \left(\log_2 \text{SNR} + \frac{-n_R \log_2 \text{SNR} + \frac{1}{\theta TB} \log_e \mathbb{E}\left\{ e^{-\theta TB \log_2 \det\left(\frac{1}{\text{SNR}}\mathbf{I} + \frac{n_R}{n_T}\mathbf{H}\mathbf{H}^\dagger\right)} \right\}}{n_R} \right) \tag{117}$$

$$= \lim_{\text{SNR}\to\infty} \frac{1}{\theta TBn_R} \log_e \mathbb{E}\left\{ e^{-\theta TB \log_2 \det\left(\frac{1}{\text{SNR}}\mathbf{I} + \frac{n_R}{n_T}\mathbf{H}\mathbf{H}^\dagger\right)} \right\} \tag{118}$$

$$= \frac{1}{\theta TBn_R} \log_e \mathbb{E}\left\{ e^{-\theta TB \log_2 \det\left(\frac{n_R}{n_T}\mathbf{H}\mathbf{H}^\dagger\right)} \right\} \tag{119}$$

$$= \log_2 \frac{n_T}{n_R} + \frac{1}{\theta TBn_R} \log_e \mathbb{E}\left\{ e^{-\theta TB \log_2 \det \mathbf{H}\mathbf{H}^\dagger} \right\}. \tag{120}$$

Above, while (116) is obtained by noting that

$$\theta TB \log_2 \det\left(\mathbf{I} + \frac{n_R}{n_T}\text{SNR}\mathbf{H}\mathbf{H}^\dagger\right) = \theta TB \log_2 \text{SNR}^{n_R} + \theta TB \log_2 \det\left(\frac{1}{\text{SNR}}\mathbf{I} + \frac{n_R}{n_T}\mathbf{H}\mathbf{H}^\dagger\right),$$

the remaining steps follow through straightforward algebraic operations. The result for the case in which $n_T < n_R$ can be readily proved by applying the above procedure to

$$\mathcal{L}_\infty = \lim_{\text{SNR}\to\infty} \left(\log_2 \text{SNR} + \frac{\frac{1}{\theta TB} \log_e \mathbb{E}\left\{ e^{-\theta TB \log_2 \det\left(\mathbf{I} + \frac{n_R}{n_T}\text{SNR}\,\mathbf{H}^\dagger\mathbf{H}\right)} \right\}}{n_T} \right). \tag{121}$$

■

Remark 7: When ergodic Shannon rate (or equivalently effective rate with $\theta = 0$) is considered, it is well-known that the high-SNR slope for the i.i.d. Rayleigh fading channel is $\mathcal{S}_\infty = \min(n_R, n_T)$. The above result shows that the high-SNR slope does not get affected by the queueing constraints when $\theta < \frac{\max(n_R,n_T)-\min(n_R,n_T)+1}{TB \log_2 e}$.

Remark 8: For the case of $\theta = 0$, it is shown in [4, Appendix B] that the power offset in the i.i.d. Rayleigh fading is[7]

$$\mathcal{L}_\infty = \begin{cases} \log_2 \frac{n_T}{n_R} - \frac{1}{n_R} E\left\{\log_2 \det \mathbf{H}\mathbf{H}^\dagger\right\} & n_R \le n_T \\ \log_2 \frac{n_T}{n_R} - \frac{1}{n_T} E\left\{\log_2 \det \mathbf{H}^\dagger\mathbf{H}\right\} & n_R > n_T \end{cases}. \tag{122}$$

By Jensen's inequality and strict concavity of the logarithm function, we have

$$\frac{1}{\theta TBn_R} \log_e E\left\{ e^{-\theta TB \log_2 \det \mathbf{H}\mathbf{H}^\dagger} \right\} > \frac{1}{\theta TBn_R} E\left\{ \log_e e^{-\theta TB \log_2 \det \mathbf{H}\mathbf{H}^\dagger} \right\} \tag{123}$$

$$= -\frac{1}{n_R} E\left\{\log_2 \det \mathbf{H}\mathbf{H}^\dagger\right\}, \quad \text{for} \quad \theta > 0 \tag{124}$$

[7] In [4], signal-to-noise ratio is defined as SNR $= \frac{n_R \mathbb{E}\{\|\mathbf{x}\|^2\}}{\mathbb{E}\{\|\mathbf{n}\|^2\}}$. Due to the presence of n_R in the numerator in the SNR definition, the first term of $\mathcal{L}_\infty$ in [4] is $\log_2 n_T$ instead of $\log_2 \frac{n_T}{n_R}$.

which shows from the comparison of (107) and (122) that the presence of queueing constraints result in higher power offset values in the high-SNR regime.

Remark 9: Note that by Hölder's inequality, we have

$$(E\{|x|^r\})^{1/r} \leq (E\{|x|^s\})^{1/s} \tag{125}$$

for $0 < r < s$. Note further that the second term in the expression of $\mathcal{L}_\infty$ can be expressed as [8]

$$\frac{1}{\theta T B n_R} \log_e E\left\{e^{-\theta T B \log_2 \det \mathbf{H}\mathbf{H}^\dagger}\right\} = \frac{1}{n_R} \log_e \left(E\left\{e^{-\theta T B \log_2 \det \mathbf{H}\mathbf{H}^\dagger}\right\}\right)^{\frac{1}{\theta T B}}. \tag{126}$$

Application of the inequality in (125) to $\left(E\left\{e^{-\theta T B \log_2 \det \mathbf{H}\mathbf{H}^\dagger}\right\}\right)^{\frac{1}{\theta T B}}$ shows that the power offset $\mathcal{L}_\infty$ in a non-decreasing function of the QoS exponent θ.

Theorem 5 characterizes $\mathcal{S}_\infty$ and $\mathcal{L}_\infty$ for a certain range of values of θ. The next result gives a partial answer to what is expected when $\theta > \frac{\max(n_R,n_T)-\min(n_R,n_T)+1}{TB\log_2 e}$, by considering the case of single-antenna transmission and reception, i.e., $n_T = n_R = 1$.

Theorem 6: In a Rayleigh fading channel with single transmit antenna and single receive antenna (i.e., $n_T = n_R = 1$), the high-SNR slope is

$$\mathcal{S}_\infty = \frac{1}{\theta T B \log_2 e} \tag{127}$$

when $\theta > \frac{1}{TB\log_2 e}$.

Proof: When we have $n_T = n_R = 1$, the effective rate expression is

$$\begin{aligned}
R_E(\text{SNR}) &= -\frac{1}{\theta T B} \log_e \mathbb{E}\left\{e^{-\theta T B \log_2\left(1+\text{SNR}|h|^2\right)}\right\} && (128)\\
&= -\frac{1}{\theta T B} \log_e \mathbb{E}\left\{e^{\log_e\left(1+\text{SNR}|h|^2\right)^{-\theta T B \log_2 e}}\right\} && (129)\\
&= -\frac{1}{\theta T B} \log_e \mathbb{E}\left\{\left(1+\text{SNR}|h|^2\right)^{-\theta T B \log_2 e}\right\} && (130)\\
&= -\frac{1}{\theta T B} \log_e \int_0^\infty (1+\text{SNR}z)^{-\theta T B \log_2 e} e^{-z}\, dz && (131)
\end{aligned}$$

where (131) follows from our Rayleigh fading assumption which implies that $z = |h|^2$ has an exponential distribution. Note that this effective rate expression can also be immediately seen to be a special case

[8] Without loss of generality, we consider the case in which $n_R \leq n_T$.

of the expressions in (100) and (101). Now, we prove the result through the following steps:

$$\begin{aligned}
\mathcal{S}_\infty &= \lim_{\text{SNR}\to\infty} \frac{R_E(\text{SNR})}{\log_2 \text{SNR}} && (132)\\
&= \lim_{\text{SNR}\to\infty} \frac{-\frac{1}{\theta TB}\log_e \int_0^\infty (1+\text{SNR}z)^{-\theta TB\log_2 e}\, e^{-z}\,dz}{\log_2 \text{SNR}} && (133)\\
&= \lim_{\text{SNR}\to\infty} \frac{-\frac{1}{\theta TB}\log_e \left[\frac{\text{SNR}}{\text{SNR}} \int_0^\infty (1+\text{SNR}z)^{-\theta TB\log_2 e}\, e^{-z}\,dz\right]}{\log_2 \text{SNR}} && (134)\\
&= \lim_{\text{SNR}\to\infty} \frac{\frac{1}{\theta TB}\log_e \text{SNR} - \frac{1}{\theta TB}\log_e \left[\text{SNR} \int_0^\infty (1+\text{SNR}z)^{-\theta TB\log_2 e}\, e^{-z}\,dz\right]}{\log_2 \text{SNR}} && (135)\\
&= \lim_{\text{SNR}\to\infty} \frac{\frac{1}{\theta TB}\log_e \text{SNR}}{\log_2 \text{SNR}} + \frac{-\frac{1}{\theta TB}\log_e \left[\text{SNR} \int_0^\infty (1+\text{SNR}z)^{-\theta TB\log_2 e}\, e^{-z}\,dz\right]}{\log_2 \text{SNR}} && (136)\\
&= \frac{1}{\theta TB\log_2 e} + \lim_{\text{SNR}\to\infty} \frac{-\frac{1}{\theta TB}\log_e \left[\text{SNR}^{-\theta TB\log_2 e+1} \int_0^\infty \left(\frac{1}{\text{SNR}}+z\right)^{-\theta TB\log_2 e} e^{-z}\,dz\right]}{\log_2 \text{SNR}} && (137)\\
&= \frac{1}{\theta TB\log_2 e} + \lim_{\text{SNR}\to\infty} \frac{-\frac{1}{\theta TB}\log_e \left[\text{SNR}^{-\theta TB\log_2 e+1}\, e^{\frac{1}{\text{SNR}}}\, \Gamma\left(-\theta TB\log_2 e+1, \frac{1}{\text{SNR}}\right)\right]}{\log_2 \text{SNR}} && (138)\\
&= \frac{1}{\theta TB\log_2 e} + \lim_{\text{SNR}\to\infty} \frac{-\frac{1}{\theta TB}\log_e \left[e^{\frac{1}{\text{SNR}}}\, \frac{\Gamma\left(-\theta TB\log_2 e+1, \frac{1}{\text{SNR}}\right)}{\frac{1}{\text{SNR}}^{-\theta TB\log_2 e+1}}\right]}{\log_2 \text{SNR}} && (139)\\
&= \frac{1}{\theta TB\log_2 e}. && (140)
\end{aligned}$$

Above, (135) is obtained by multiplying the integral inside the logarithm in the numerator by $\frac{\text{SNR}}{\text{SNR}}$ as shown in (134). and by using the fact that the logarithm of the division is equal to the difference of the logarithms. (136) follows by separately writing the fractions. (137) is obtained by evaluating the limit of the first fraction, and by expressing $(1+\text{SNR}z)^{-\theta TB\log_2 e}$ in the second fraction as $\text{SNR}^{-\theta TB\log_2 e}\left(\frac{1}{\text{SNR}}+z\right)^{-\theta TB\log_2 e}$. (138) follows from the fact that [33, Equation 3.382.4]

$$\int_0^\infty \left(\frac{1}{\text{SNR}}+z\right)^{-\theta TB\log_2 e} e^{-z}\,dz = e^{\frac{1}{\text{SNR}}}\,\Gamma\left(-\theta TB\log_2 e, \frac{1}{\text{SNR}}\right) \qquad (141)$$

where $\Gamma(\alpha, x)$ is the upper incomplete Gamma function. (139) is obtained by rearranging the terms in the numerator of the fraction in the second term. Finally, (140) follows by realizing that the limiting expression in (139) is equal to zero. This is noted from the fact that as $\text{SNR}\to\infty$, we have

$$e^{\frac{1}{\text{SNR}}} \longrightarrow 1 \qquad (142)$$

$$\frac{\Gamma\left(-\theta TB\log_2 e+1, \frac{1}{\text{SNR}}\right)}{\frac{1}{\text{SNR}}^{-\theta TB\log_2 e+1}} \longrightarrow \frac{1}{\theta TB\log_2 e - 1}, \qquad (143)$$

indicating that the numerator in the limiting expression in (139) is approaching a finite value as SNR increases while the denominator grows without bound. The limit in (143) is due to the fact that [9]

$$\frac{\Gamma(\alpha, x)}{x^{\alpha}} \to \frac{-1}{\alpha} \quad \text{as} \quad x \to 0 \tag{144}$$

when $\alpha < 0$, which is satisfied in our setting from our assumption that $\theta TB \log_2 e > 1$. ■

Remark 10: Theorem 6 shows for the single-antenna case that when $\theta > \frac{\max(n_R,n_T)-\min(n_R,n_T)+1}{TB\log_2 e} = \frac{1}{TB\log_2 e}$, the high-SNR slope is $\mathcal{S}_\infty = \frac{1}{\theta TB\log_2 e} < \min(n_R, n_T) = 1$, and diminishes with increasing θ. Note that by Theorem 5, $\mathcal{S}_\infty = \min(n_R, n_T) = 1$ when $\theta < \frac{1}{TB\log_2 e}$ in the case of single antennas at the receiver and transmitter.

Remark 11: For the multiple-antenna case, we have the following additional discussion. An expression for the components of the Hankel matrix $\mathbf{G}$ is given by [28]

$$g_{i,j} = \int_0^\infty \left(1 + \frac{n_R}{n_T}\,\text{SNR}\,z\right)^{-\theta TB\log_2 e} z^{i+j+d}\, e^{-z}\, dz \qquad i,j = 0,1,\ldots,k-1 \tag{145}$$

$$\begin{aligned} &= \frac{\pi}{\Gamma\left(\theta TB\log_2 e\right)\sin\left(\pi\left(d+i+j-\theta TB\log_2 e\right)\right)}\times \\ &\left[\frac{\left(\frac{n_R}{n_T}\,\text{SNR}\right)^{-1-d-i-j}\Gamma(1+d+i+j)}{\Gamma\left(2+d+i+j-\theta TB\log_2 e\right)}\,{}_1F_1\left(1+d+i+j, 2+d+i+j-\theta TB\log_2 e, \frac{n_T}{n_R\text{SNR}}\right)\right. \\ &\left. - \frac{\left(\frac{n_R}{n_T}\,\text{SNR}\right)^{-\theta TB\log_2 e}\Gamma(\theta TB\log_2 e)}{\Gamma\left(-d-i-j+\theta TB\log_2 e\right)}\,{}_1F_1\left(\theta TB\log_2 e, -d-i-j+\theta TB\log_2 e, \frac{n_T}{n_R\text{SNR}}\right)\right] \end{aligned} \tag{146}$$

where ${}_1F_1$ denotes the confluent hypergeometric function and has the following series expansion [33]

$${}_1F_1(a,b,z) = \sum_{i=0}^{\infty}\frac{(a)_i z^i}{(b)_i i!} = 1 + \frac{a}{b}\frac{z}{1!} + \frac{a(a+1)}{b(b+1)}\frac{z^2}{2!} + \frac{a(a+1)(a+2)}{b(b+1)(b+2)}\frac{z^3}{3!} + \ldots \tag{147}$$

Note that the expression in (146) is valid when $\theta TB\log_2 e \neq \pm(d+i+j)$ for all i, j because of the presence of the sinousoid in the denominator of the first term and the fact that $\Gamma(x) = \infty$ or $-\infty$ when x is a negative integer. Under this restriction, we can see (by also noting that ${}_1F_1(a,b,0) = 1$) that the first term inside the square brackets in (146) scales as $\text{SNR}^{-1-d-i-j}$ while the second term scales as $\text{SNR}^{-\theta TB\log_2 e}$ as SNR $\to \infty$. Note that $d = \max(n_R, n_T) - \min(n_R, n_T)$ and $i, j = 0, 1, \ldots, \min(n_R, n_T) -$

[9] The limit in (144) can be obtained from the following facts: A definition of the upper incomplete Gamma function is given by [33, Equation 8.351.4] $\Gamma(\alpha, x) = x^\alpha e^{-x}\Psi(1, 1+\alpha; x) = x^\alpha e^{-x}\int_0^\infty e^{-xt}(1+t)^{\alpha-1}dt$. From this definition, we can easily see that $\lim_{x\to 0}\frac{\Gamma(\alpha,x)}{x^\alpha} = \int_0^\infty (1+t)^{\alpha-1}dt = \frac{-1}{\alpha}$ for $\alpha < 0$.

1. Therefore, when

$$\theta TB\log_2 e > 1 + d + 2(\min(n_R, n_T) - 1) = \max(n_R, n_T) + \min(n_R, n_T) - 1, \tag{148}$$

the first terms with $\text{SNR}^{-1-d-i-j}$ will dictate the rate at which $g_{i,j}$'s approach zero for all i, j. Hence, we have

$$g_{i,j} \sim \frac{\pi}{\Gamma\left(\theta TB\log_2 e\right)\sin\left(\pi\left(d+i+j-\theta TB\log_2 e\right)\right)} \frac{\left(\frac{n_R}{n_T}\text{SNR}\right)^{-1-d-i-j}\Gamma(1+d+i+j)}{\Gamma\left(2+d+i+j-\theta TB\log_2 e\right)} \tag{149}$$

as $\text{SNR} \to \infty$. Note that the matrix $\widetilde{\mathbf{G}}$, whose components $\tilde{g}_{i,j}$ are equal to the right-hand side of (149), is still a Hankel matrix as the components depend on the indexes only through $(i+j)$. If the determinant is nonzero, it can be easily verified that the determinant of $\widetilde{\mathbf{G}}$ scales as

$$\det(\widetilde{\mathbf{G}}) \sim \text{SNR}^{-\sum_{i=1}^{\min(n_R,n_T)}(2i-1)} = \text{SNR}^{(\min(n_R,n_T))^2}. \tag{150}$$

For instance,

$$\det(\widetilde{\mathbf{G}}) = \det\left(\begin{bmatrix} a\text{SNR}^{-1} & b\text{SNR}^{-2} & c\text{SNR}^{-3} \\ b\text{SNR}^{-2} & c\text{SNR}^{-3} & d\text{SNR}^{-4} \\ c\text{SNR}^{-3} & d\text{SNR}^{-4} & e\text{SNR}^{-5} \end{bmatrix}\right) \sim \text{SNR}^{-(1+3+5)} = \text{SNR}^{-9} \tag{151}$$

for large SNR as long as the constant $a, b, c, d,$ and e are such that $\det(\widetilde{\mathbf{G}})$ is nonzero. Finally, we have under the aforementioned conditions that

$$R_{E,\text{id}}(\text{SNR}) \sim -\frac{1}{\theta TB}\log_e\left(\frac{\det(\widetilde{\mathbf{G}}(\theta,\text{SNR}))}{\prod_{i=1}^{k}\Gamma(d+i)}\right) \sim \frac{(\min(n_R,n_T))^2}{\theta TB\log_2 e}\log_2 \text{SNR}, \tag{152}$$

indicating that

$$\mathcal{S}_\infty = \frac{(\min(n_R,n_T))^2}{\theta TB\log_2 e} \tag{153}$$

when $\theta TB\log_2 e > \max(n_R, n_T) + \min(n_R, n_T) - 1$. Note that under this condition on θ, $\mathcal{S}_\infty = \frac{\min(n_R,n_T)^2}{\theta TB\log_2 e} < \min(n_R, n_T)$. Note also that the above conclusion reduces to the result of Theorem 6 when $n_R = n_T = 1$.

VII. Numerical Results

In this section, we numerically illustrate the analytical results obtained in the previous sections. In order to treat the low-SNR and high-SNR regimes jointly, we consider the i.i.d. Rayleigh fading channel in

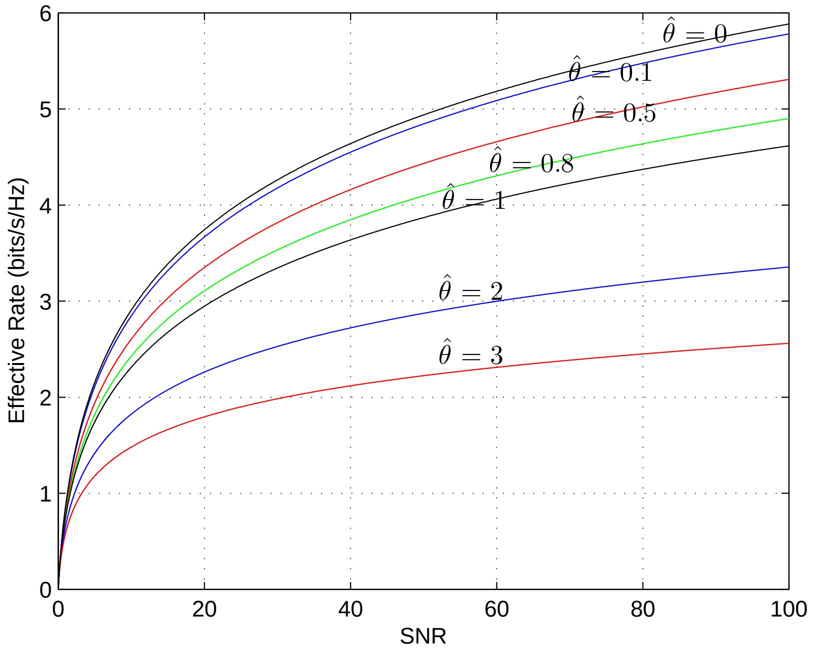

Fig. 1. Effective rate R_E vs. SNR in the single-antenna case (i.e., when $n_R = n_T = 1$) for different values of $\hat{\theta} = \theta T B \log_2 e$.

which the components of the channel matrix $\mathbf{H}$ are i.i.d. zero-mean, unit-variance, circularly symmetric Gaussian random variables. We further assume that the input covariance matrix is $\mathbf{K}_x = \frac{1}{n_T}\mathbf{I}$, and the effective rate is given by

$$R_{E,\text{id}}(\text{SNR}) = -\frac{1}{\theta T B} \log_e \mathbb{E}\left\{\exp\left(-\theta T B \log_2 \det\left(\mathbf{I} + \frac{n_R}{n_T}\text{SNR}\,\mathbf{H}\mathbf{H}^\dagger\right)\right)\right\} \text{ bits/s/Hz.} \tag{154}$$

Under these assumptions, we can easily compute the effective rate by using the formulation in (101) and performing integral computations. We note that the computations of the effective rate in the correlated fading case can be done using the expressions of the moment generating function of the mutual information of correlated MIMO Gaussian fading channels provided in [29]. Summary of such non-asymptotic results, along with asymptotic spectrum theorems, on random matrices is presented in [30].

Figure 1 plots the effective rate $R_{E,\text{id}}$ as a function of SNR in the single-antenna case ($n_R = n_T = 1$) for different values of $\hat{\theta} = \theta T B \log_2 e$. It is assumed that $T = 1$ ms $= 10^{-3}$ s and $B = 100$kHz $= 10^5$Hz. Note that when $\hat{\theta} = 0$ or equivalently $\theta = 0$, there are no statistical queueing constraints and the effective capacity is equal to the ergodic Shannon capacity. In Fig. 1, we observe that the effective rate in general diminishes with increasingly more strict queueing constraints (or equivalently higher θ values). As expected, under more strict buffer constraints, lower arrival rates are supported, and as

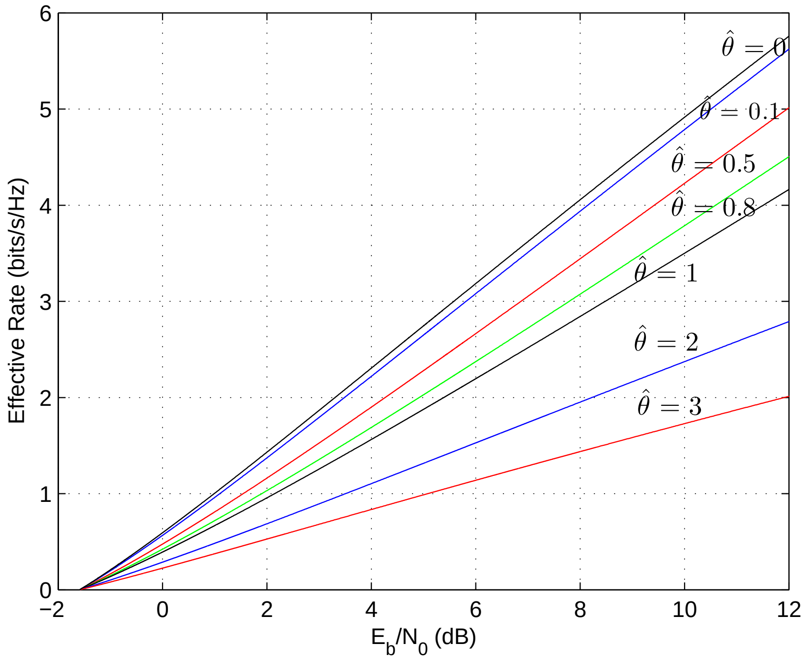

Fig. 2. Effective rate R_E vs. bit energy $\frac{E_b}{N_0}$ in the single-antenna case (i.e., when $n_R = n_T = 1$) for different values of $\hat{\theta} = \theta T B \log_2 e$.

a result, lower departure rates are seen. On the other hand, as predicted by the low-SNR results of Section IV, all rate curves have the same slope at SNR $= 0$. Note that this slope is the one achieved in the absence of QoS constraints (i.e., when $\theta = 0$). Therefore, the impact of queueing constraints on the performance lessens at low SNR values. An intuitive explanation of this observation is that as power decreases, arrival rates that can be supported by the system diminishes as well, which in turn decreases the effect of buffer violation constraints. Note also that as discussed in Section V, results similar to those in the low-power regime are obtained in the wideband regime if the channel experiences rich multipath fading. Therefore, another interpretation of the above observation is that QoS constraints have less impact on the performance as the bandwidth increases in rich multipath environments. This is due to the fact that the number of noninteracting subchannels and hence the number of degrees of freedom increases with increasing bandwidth, and the system has increasingly higher diversity to combat with buffer constraints.

Fig. 1 confirms the analytical high-SNR results as well. As predicted by Theorem 5, the high-SNR slope is the same as that achieved in the absence of QoS constraints as long as $\hat{\theta} = \theta T B \log_2 e < 1$. On the other hand, as proved in Theorem 6, high-SNR slope is strictly less than 1 when $\hat{\theta} > 1$. The difference in the rates of increase at high SNRs is clearly seen in Fig. 1.

In Fig. 2, we plot the effective rate as a function of the bit energy in the single-antenna case.

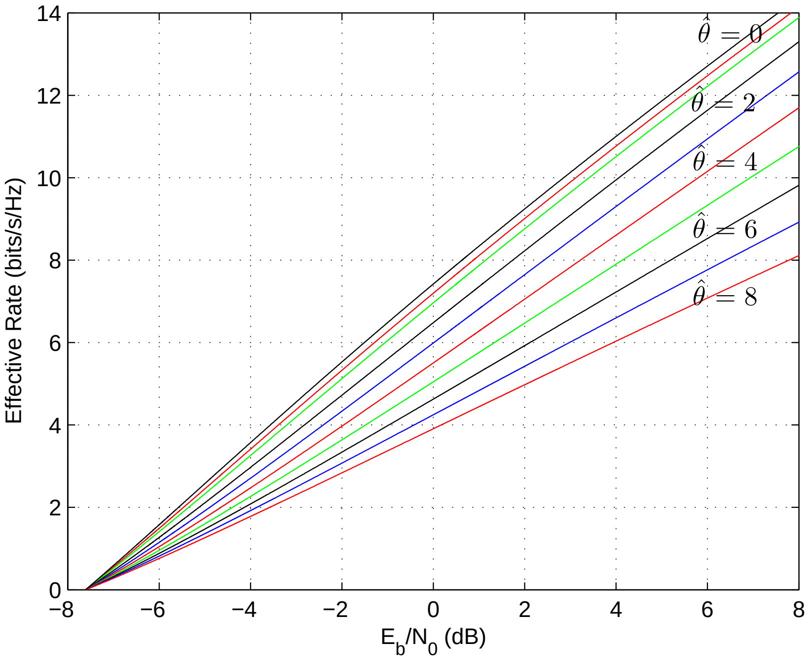

Fig. 3. Effective rate R_E vs. bit energy $\frac{E_b}{N_0}$ for $\hat{\theta} = \theta TB \log_2 e = 0, 0.5, 1, 2, 3, 4, 5, 6, 7, 8$ when $n_R = 2$ and $n_T = 5$.

Confirming the discussion in Section IV-B, we immediately note that the minimum bit energy for all values of θ is -1.59 dB, which is the fundamental limit in the absence of QoS limitations. This is a consequence of the fact that the effective rate curves as a function of SNR have the same slope at zero SNR. However, since the second derivatives of the effective rate at $\text{SNR} = 0$ decreases with increasing θ, we observe in Fig. 2 that we have smaller wideband slopes, $\mathcal{S}_0$, for larger values of θ. Similarly as in Fig. 1, we observe smaller high-SNR slopes, $\mathcal{S}_\infty$, when $\hat{\theta} > 1$.

In Fig. 3, effective rate vs. bit energy curves are plotted under the assumption that the number of receive antennas is $n_R = 2$ and the number of transmit antennas is $n_T = 5$. We still assume that $T = 1$ ms and $B = 100$kHz. In the figure, the curves from the top to the bottom are for $\hat{\theta} = 0, 0.5, 1, 2, 3, 4, 5, 6, 7, 8$ in this order [10]. We again immediately note that the same minimum bit energy is attained for all values of $\hat{\theta}$ while the wideband slopes $\mathcal{S}_0$ are smaller for larger values of the QoS exponent. In this case, the minimum bit energy is $\frac{E_b}{N_0}_{\min} = 10 \log_{10} \left(\frac{\log_e 2}{n_R^2} \right) = -7.61$ dB [11]. At high SNR levels, we observe that, as shown in Theorem 5, when $\hat{\theta} = \theta TB \log_2 e < \max(n_R, n_T) - \min(n_R, n_T) + 1 = 4$, $\mathcal{S}_\infty$ is the same as that achieved when $\hat{\theta} = 0$ (i.e., when $\theta = 0$). For $\hat{\theta} > 4$, we note the gradual decrease in the high-SNR

[10] Note that when $\theta = 0$, effective capacity becomes equal to the ergodic Shannon capacity. For this case, rate is computed using the formulation provided in [25, Theorem 2].

[11] As opposed to (64) where $\frac{E_b}{N_0}_{\min} = \frac{\log_e 2}{n_R}$, we have $\frac{E_b}{N_0}_{\min} = \frac{\log_e 2}{n_R^2}$ in the figure since we plot the effective rate in bits/s/Hz without normalization with the number of receive antennas.

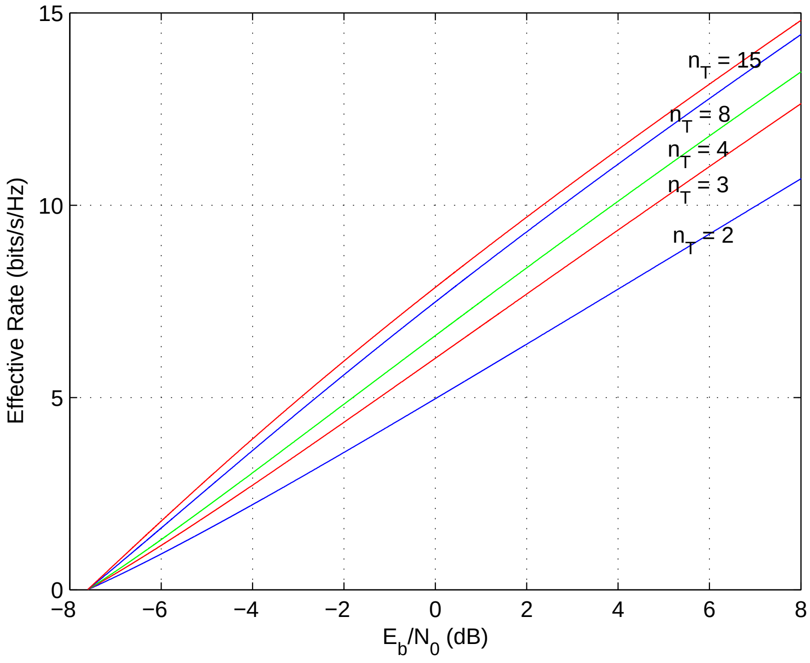

Fig. 4. Effective rate R_E vs. bit energy $\frac{E_b}{N_0}$ for $n_T = 2, 3, 4, 8, 15$ when $n_R = 2$ and $\hat{\theta} = \theta T B \log_2 e = 1$.

slope.

When we compare Figs. 2 and 3, we see that the rate curves are much closer to each other in Fig. 3, indicating the resilience provided by spatial diversity against queueing constraints. This is further illustrated in Fig. 4, where effective rate vs. bit energy curves are plotted for different number of transmit antennas when $n_R = 2$ and $\hat{\theta} = 1$. In this figure, we observe that the wideband slope $\mathcal{S}_0$ increases with increasing number of transmit antennas for a given QoS exponent θ. Moreover, we note that improvements are provided at all SNR levels when the number of antennas is increased in the system, again pointing to the benefits of spatial diversity.

Heretofore, the discussions on the low-SNR regime apply to the cases in which the transmit power is small or the bandwidth is large but in a rich multipath fading setting. In Section V, we have remarked that sparse multipath fading has considerable impact on the performance in the wideband regime. In order to numerically illustrate these results, we provide Figs. 5 and 6. In Fig. 5, effective rate

$$R_{E,\text{id}} = -\frac{1}{\theta T B_c} \log_e \mathbb{E}\left\{\exp\left(-\theta T B_c \log_2 \det\left(\mathbf{I} + \frac{n_R}{n_T}\text{SNR}\mathbf{H}\mathbf{H}^\dagger\right)\right)\right\} \tag{155}$$

is plotted as a function of the bit energy. Above in (155), B_c denotes the coherence bandwidth, and $\text{SNR} = \frac{P}{n_R m B_c N_0}$ where m is the number of noninteracting subchannels, each experiencing i.i.d. zero-mean, unit-variance Gaussian fading. In this figure, we have $n_R = n_T = 2$, and $\frac{P}{N_0} = 10^4$, $T = 1$ms.

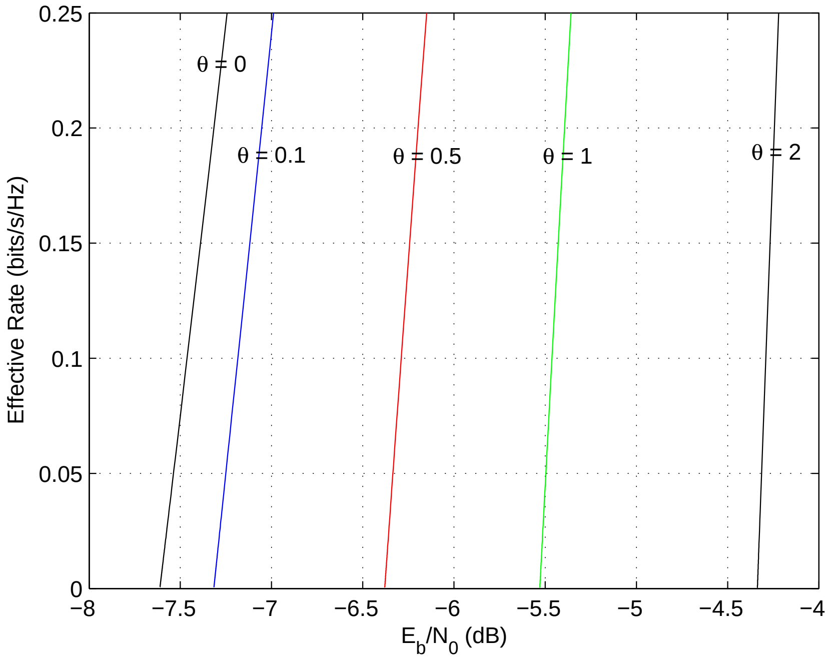

Fig. 5. Effective rate R_E vs. bit energy $\frac{E_b}{N_0}$ for $\theta = 0, 0.1, 0.5, 1, 2$ in sparse wideband fading channels. $n_R = n_T = 2$. The number of subchannels is $m = 5$. The coherence bandwidth B_c increases with increasing bandwidth.

We consider the setting in which the number of subchannels is bounded while the coherence bandwidth increases with increasing bandwidth. We assume $m = 5$ and plot the curves by varying B_c from 10kHz to 10MHz. Therefore, bandwidth increases from 50kHz to 50MHz. As predicted by the result of Theorem 3, the minimum bit energy depends on θ and increases with increasing θ. We note that for relatively large values of θ, considerably higher bit energies are needed when compared with the case of $\theta = 0$.

In Fig. 5, we have assumed that the number of subchannels and hence the number of degrees of freedom is bounded, and B_c increases linearly with increasing bandwidth. We have seen that having bounded number of degrees of freedom induces substantial energy penalty especially if the queueing constraints are stringent. Another scenario in sparse multipath fading is the one in which B_c increases but only sublinearly with B. In such a case, the number of subchannels m increases with B as well. In Theorem 4, we have shown for this scenario that the same minimum bit energy as in the case of $\theta = 0$ can be attained. This is depicted in Fig. 6. In this figure, the parameters are the same as in Fig. 5, except we now assume that m increases from 5 to 100 as B_c increases from 10kHz to 10MHz. We note that in all cases, the minimum bit energy of -7.61 dB is approached. However, it is interesting to observe that the wideband slopes are zero when $\theta > 0$, indicating that approaching the minimum bit energy is very demanding in terms of bandwidth in the presence of queueing constraints.

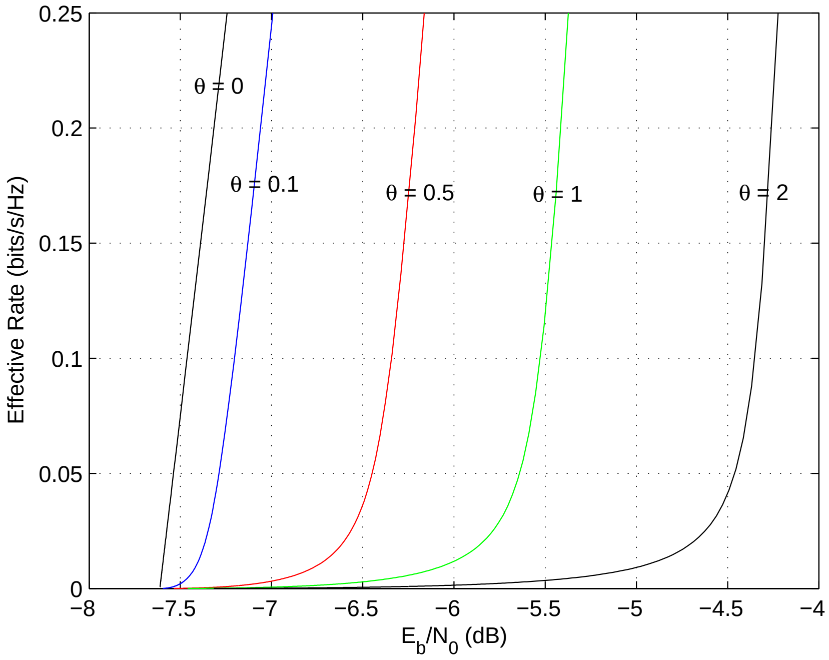

Fig. 6. Effective rate R_E vs. bit energy $\frac{E_b}{N_0}$ for $\theta = 0, 0.1, 0.5, 1, 2$ in sparse wideband fading channels. $n_R = n_T = 2$. Both the coherence bandwidth B_c and the number of subchannels m increase with increasing bandwidth.

VIII. Conclusion

In this paper, we have investigated the performance of MIMO wireless systems operating under statistical queueing (or QoS) constraints, which are formulated as limitations on buffer violation probabilities in the large-queue-length regime. We have employed effective capacity as the performance metric that provides the throughput under such constraints. We have studied the effective capacity in the low-power, wideband, and high-SNR regimes. In the low-power regime, we have obtained expressions for the first and second derivatives of the effective capacity at zero SNR under various assumptions on the channel knowledge at the transmitter side. We have shown that while the first derivative does not depend on the QoS constraints, the second derivative diminishes as these constraints become more stringent. As a byproduct of these results, we have demonstrated that the minimum bit energy requirements in the presence of QoS constraints in the low-power regime are the same as those required in the absence of such constraints. However, the wideband slope is shown to significantly get affected by queueing constraints.

Results derived in the low-power regime are proven to apply to the wideband regime in rich multipath fading environments. On the other hand, we have noted that sparse multipath fading induces energy penalty if the number of noninteracting subchannels remains bounded in the wideband regime. In this case, the minimum bit energy is shown to depend on the QoS exponent θ. If the number of subchannels increase with bandwidth but only sublinearly, we have seen that the minimum bit energy required in the absence of buffer constraints can be attained, but we have demonstrated in the numerical results that approaching this level is very slow.

Finally, we have investigated the performance in the high-SNR regime by determining the high-SNR slope and power offset values. In particular, we have shown that if the QoS exponent is less than a certain thereshold, the high-SNR slope of $\min(n_R, n_T)$ can be maintained. However, in this case, we have remarked that there is still a price to be paid in terms of the power offset $\mathcal{L}_\infty$ when queueing limitations are present. For the single-antenna case, we have proven that increasing θ beyond a threshold starts affecting the high-SNR slope $\mathcal{S}_\infty$. In such a case, $\mathcal{S}_\infty$ is shown to diminish with increasing θ. We have discussed extensions of this result to the multiple-antenna scenarios, and illustrated them through numerical results.

References

[1] A. Goldsmith, S. A. Jafar, N. Jindal, and S. Vishwanath, “Capacity limits of MIMO channels,” *IEEE Journ. on Select. Areas in Commun.*, vol. 21, no. 5, pp.684702, June 2003.

[2] S. Verdú, “Spectral efficiency in the wideband regime,” *IEEE Trans. Inform. Theory*, vol.48, no.6, pp.1319-1343, Jun. 2002.

[3] A. Lozano, A. M. Tulino, and S. Verdú, “Multiple-antenna capacity in the low-power regime,” *IEEE Trans. Inform. Theory*, vol.49, no.10, pp.2527-2544, Oct. 2003.

[4] A. Lozano, A. M. Tulino, and S. Verdú, “High-SNR Power Offset in Multiantenna Communication,” *IEEE Trans. Inform. Theory*, vol.51, no.12 pp.4134–4151, Dec. 2005.

[5] D. Wu and R. Negi, “Effective capacity: a wireless link model for support of quality of service,” *IEEE Trans. Wireless Commun.*, vol.2,no. 4, pp.630-643. July 2003

[6] C.-S. Chang, “Stability, queue length, and delay of deterministic and stochastic queuing networks,” *IEEE Trans. Auto. Control*, vol. 39, no. 5, pp. 913-931, May 1994

[7] C.-S. Chang, *Performance Guarantees in Communication Networks*, New York: Springer, 1995

[8] C.-S. Chang and T. Zajic, “Effective bandwidths of departure processes from queues with time varying capacities,” *Proceedings of IEEE Infocom*, pp. 1001-1009, 1995

[9] J. Tang and X. Zhang, “Quality-of-service driven power and rate adaptation over wireless links,” *IEEE Trans. Wireless Commun.*, vol. 6, no. 8, pp.3058-3068, Aug. 2007.

[10] J. Tang and X. Zhang, “Quality-of-service driven power and rate adaptation for multichannel communications over wireless links,” *IEEE Trans. Wireless Commun.*, vol. 6, no. 12, pp.4349-4360, Dec. 2007.

[11] J. Tang and X. Zhang, “Cross-layer-model based adaptive resource allocation for statistical QoS guarantees in mobile wireless networks,” *IEEE Trans. Wireless Commun.*, vol. 7, pp.2318-2328, June 2008.

[12] L. Liu, P. Parag, J. Tang, W.-Y. Chen and J.-F. Chamberland, “Resource allocation and quality of service evaluation for wireless communication systems using fluid models,” *IEEE Trans. Inform. Theory*, vol. 53, no. 5, pp. 1767-1777, May 2007

[13] L. Liu, P. Parag, and J.-F. Chamberland, “Quality of service analysis for wireless user-cooperation networks,” *IEEE Trans. Inform. Theory*, vol. 53, no. 10, pp. 3833-3842, Oct. 2007

[14] M.C. Gursoy, D. Qiao, and S. Velipasalar, "Analysis of energy efficiency in fading channel under QoS constrains," *IEEE Trans. Wireless Commun.*, vol. 8, no. 8, pp. 4252-4263, Aug. 2009.

[15] D. Qiao, M.C. Gursoy, and S. Velipasalar, "The impact of QoS constraints on the energy efficiency of fixed-rate wireless transmissions," accepted for publication in the *IEEE Trans. Wireless Commun.*, 2009; conference version appeared at the IEEE International Conference on Communications (ICC), Jun. 2009.

[16] D. Qiao, M.C. Gursoy, and S. Velipasalar, "Energy efficiency of fixed-rate wireless transmissions under queueing constraints and channel uncertainty," to appear at the at the IEEE Global Communications Conference (GLOBECOM), Dec. 2009.

[17] E. A. Jorswieck, R. Mochaourab, and M. Mittelbach, "Effective capacity maximization in multi-antenna channels with covariance feedback," IEEE Intenational Conference on Communications, Dresden, Germany, 2009.

[18] L. Liu, and J.-F. Chamberland, "On the effective capacities of multiple-antenna Gaussian channels," IEEE International Symposium on Information Theory, Toronto, 2008.

[19] M. Luby, "LT codes," Proc. 43rd Ann. IEEE Symp. Found. Comp. Sci., 2002, pp. 271280.

[20] A. Shokrollahi, "Raptor codes," *IEEE Trans. Inform. Theory*, vol. 52, pp. 2551-2567, June 2006.

[21] J. Casture and Y. Mao, "Rateless coding and relay networks," *IEEE Signal Process. Mag.*, vol. 24, pp. 27-35, Sept. 2007.

[22] J. Casture and Y. Mao, "Rateless coding over fading channels," *IEEE Comm. Letters*, vol. 10, pp. 46-48, Jan. 2006.

[23] D. Porrat, D. N. C. Tse, and S. Nacu, "Channel uncertainty in ultra-wideband communication systems," *IEEE Trans. Inform. Theory*, vol. 53, pp. 194-208, Jan. 2007.

[24] V. Raghavan, G. Hariharan, and A. M. Sayeed, "Capacity of sparse multipath channels in the ultra-wideband regime," *IEEE Journ. Select. Topics in Signal Processing*, vol. 1, pp. 357-371, Oct. 2007.

[25] E. Telatar and D. N. C. Tse, "Capacity and mutual information of wideband multipath fading channels," *IEEE Trans. Inform. Theory*, vol. 46, pp. 1384-1400, July 2000.

[26] I. E. Telatar, "Capacity of multi-antenna Gaussian channels," *Europ. Trans. Telecomm.*, vol. 10, no. 6, pp. 585–596, Nov.-Dec. 1999.

[27] S. Shamai (Shitz) and S. Verdú, "The impact of frequency-flat fading on the spectral efficiency of CDMA," *IEEE Trans. Inform. Theory*, vol. 47, no. 4 pp. 1302–1327, May 2001.

[28] Z. Wang and G. B. Giannakis, "Outage mutual information of space-time MIMO channels," *IEEE Trans. Inform. Theory*, vol. 50, no. 4 pp. 657–662, Apr. 2004.

[29] M. Kießling "Unifying analysis of ergodic MIMO capacity in correlated Rayleigh fading environments," *European Trans. on Telecommun.* vol. 16, no. 1, pp. 17-35, Jan./Feb. 2005.

[30] A. M. Tulino and S. Verdú, "Random Matrix Theory and Wireless Communications," *Foundations and Trends in Communications and Information Theory*, vol. 1, no. 1, pp. 1–182, 2004.

[31] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, 1999.

[32] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[33] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, Academic Press, 2007.